DOCUMENT RESUME

ED 292 864	TM 011 270

AUTHOR	Peterson, Norman G., Ed.
TITLE	Development and Field Test of the Trial Battery for Project A. Improving the Selection, Classification and Utilization of Army Enlisted Personnel. Project A: Improving the Selection, Classification and Utilization of Army Enlisted Personnel. ARI Technical Report 739.
INSTITUTION	Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
PUB DATE	May 87
CONTRACT	MDA-903-82-C-0531
NOTE	413p.
AVAILABLE FROM	U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC17 Plus Postage.
DESCRIPTORS	Biographical Inventories; Classification; Cognitive Measurement; Computer Assisted Testing; *Enlisted Personnel; *Field Tests; Interest Inventories; Job Performance; Military Personnel; Occupational Tests; Perception Tests; Personality Measures; Personnel Evaluation; *Personnel Selection; Predictive Measurement; Psychomotor Skills; *Test Construction
IDENTIFIERS	*Armed Services Vocational Aptitude Battery; Paper and Pencil Tests

ABSTRACT

As part of the United States Army's Project A, research has been conducted to develop and field test a battery of experimental tests to complement the Armed Services Vocational Aptitude Battery in predicting soldiers' job performance. Project A is the United States Army's large-scale manpower effort to improve selection, classification, and utilization of enlisted personnel. Findings from an extensive literature review and associated expert judgments and administration of a preliminary battery of "off-the-shelf" measures guided the development of new tests. Three major types of tests were prepared, namely, paper-and-pencil tests of cognitive ability; computer-administered tests of perceptual and psychomotor abilities; and paper-and-pencil inventories measuring temperament, biographical data, and vocational interests. After iterative pilot tests and revisions, the measures were field tested. Analysis indicted that the new tests had adequate to excellent psychometric qualities, were relatively unique, and were not unduly affected by practice or by faking in an applicant setting. The resulting trial battery contains six cognitive tests; 10 computer-administered perceptual and psychomotor tests; and two inventories measuring temperament, biographical factors, and interests. The battery is being used in the next Project A phase, which involves concurrent validation to evaluate predictor measures against subsequent job performance. (TJH)

ED 292864

Technical Report 739

Project A:
Improving the Selection, Classification and
Utilization of Army Enlisted Personnel

Development and Field Test of the Trial Battery for Project A

Norman G. Peterson, Editor
Personnel Decisions Research Institute

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

U.S. Army
Research Institute for the Behavioral and Social Sciences

May 1987

Approved for public release; distribution unlimited.

TM 011 270

BEST COPY AVAILABLE

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON **Technical Director**	**WM. DARRYL HENDERSON** **COL, IN** **Commanding**

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Deirdre Knapp
Elizabeth Smith
Hilda Wing

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE --

3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 739

6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization

6b. OFFICE SYMBOL *(If applicable)* HumRRO

7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences

6c. ADDRESS *(City, State, and ZIP Code)*

1100 South Washington Street
Alexandria, VA 22314-4499

7b. ADDRESS *(City, State, and ZIP Code)*

5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8a. NAME OF FUNDING / SPONSORING ORGANIZATION --

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA 903-82-C-0531

8c. ADDRESS *(City, State, and ZIP Code)* --

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
6.37.31.A	2Q263-731A792	2.3.2	2.3.2.C.1

11. TITLE *(Include Security Classification)* Development and Field Test of the Trial Battery for Project A

12. PERSONAL AUTHOR(S) Norman G. Peterson, Editor (Personnel Decisions Research Institute)

13a. TYPE OF REPORT Final Report

13b. TIME COVERED FROM Oct 82 TO Sep 85

14. DATE OF REPORT *(Year, Month, Day)* 1987, May

15. PAGE COUNT 391

16. SUPPLEMENTARY NOTATION Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute).

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)* Classification, Cognitive Measures, Computer-Administered Tests, Field Tests, Non-Cognitive Measures, Predictor Measures, Project A, Psychomotor Tests, Selection, Trial Battery

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This research was performed under Project A, the U.S. Army's large-scale manpower effort to improve selection, classification, and utilization of enlisted personnel. This report deals with development and field test of a battery of experimental tests to complement the Armed Services Vocational Aptitude Battery in predicting soldiers' job performance.

Findings from an extensive literature review, expert judgments on validity of measures identified in the review, and administration of a preliminary battery of "off-the-shelf" measures guided the development of new tests. Three major types were prepared: paper-and-pencil tests of cognitive ability; computer-administered tests of perceptual/psychomotor abilities; and paper-and-pencil inventories measuring temperament, biographical data, and vocational interests. After iterative pilot tests and revisions,
(Continued)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL Lawrence M. Hanser

22b. TELEPHONE *(Include Area Code)* (202) 274-8275

22c. OFFICE SYMBOL PERI-RS

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

ARI Technical Report 739

19. Abstract (Continued)

the measures were field tested. Analysis indicated the new tests had adequate to excellent psychometric qualities, were relatively unique, and were not unduly affected by practice or by faking in an applicant setting.

The resulting Trial Battery contains six cognitive paper-and-pencil tests, 10 computer-administered perceptual/psychomotor tests, and two paper-and-pencil inventories measuring temperament, biodata, and interests. It is being used in the next Project A phase, concurrent validation executed with FY83/84 accessions to evaluate the predictor measures against subsequent job performance.

This report is supplemented by a limited-distribution Research Note, Test Appendixes to ARI Technical Report 739: Development and Field Test of the Trial Battery for Project A, ARI Research Note 87-24, April 1987.

Technical Report 739

Project A:
Improving the Selection, Classification and Utilization of Army Enlisted Personnel

Development and Field Test of the Trial Battery for Project A

Norman G. Peterson, Editor
Personnel Decisions Research Institute

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

May 1987

Army Project Number
2Q263731A792

Manpower and Personnel

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This document describes the development and field testing of a trial battery of newly constructed predictor measures for evaluating the potential performance of Army applicants. The research was part of the Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance. The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A." Another part of the effort is the development of a prototype Computerized Personnel Allocation System, referred to as "Project B." Together, these Army Research Institute research efforts, with their in-house and contract components, comprise a landmark program to develop a state-of-the-art, empirically validated personnel selection, classification, and allocation system.

Edgar M Johnson
EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

Hilda Wing of the U.S. Army Research Institute has provided invaluable assistance in many ways during the course of development of the Trial Battery. Lloyd Humphreys and Jay Uhlaner, members of the Scientific Advisory Group for Project A, have provided shrewd insights to greatly bolster the scientific rigor and practical relevance of our work. There is simply no substitute for the wisdom and experience that these two men have shared with us. John Campbell has kept our vision clearly focused on the scientific tasks at hand and contributed his excellent psychometric prowess as well. Kent Eaton has acted superbly as our Army "conscience" with his tireless quest for precision and clarity in our work. Laurie Wise and Winnie Young provided excellent and timely data analysis expertise and support at various points in this effort. Lola Zook provided invaluable assistance in revising and restructuring this report. Carol Wyman has been brilliant and tireless in her efforts to produce innumerable, readable documents from the nearly indecipherable handwriting provided her. Finally, Marv Goer has patiently tolerated, as well as supported, a group of researchers not unfairly characterized as "prima donna."

The greatest and most heart-felt acknowledgment goes, however, to all the Army personnel who contributed their time and effort to the development, pilot testing, and field testing of the measures of the Trial Battery. They are too numerous to mention, but we thank them all for the sincere effort and high motivation they showed as they worked with us on Project A.

DEVELOPMENT AND FIELD TEST OF THE TRIAL BATTERY FOR PROJECT A

EXECUTIVE SUMMARY

Requirement:

Project A is a large-scale, multiyear research program intended to improve the selection and classification for initial assignment of persons to U.S. Army Military Occupational Specialties (MOS). A comprehensive set of job performance measures are being developed to assess the validity of the Armed Services Vocational Aptitude Battery (ASVAB) and a set of newly developed experimental predictor measures.

This report describes the development and field test of the newly developed predictor measures.

Procedure:

Initial work concentrated on the development of a theoretical approach and research design to effectively and efficiently accomplish the research objective: the development of new predictor tests and inventories that would complement the Armed Services Vocational Aptitude Battery (ASVAB), primarily by measuring abilities that would be valid for predicting soldiers' job performance but were not measured on the ASVAB.

Early activities included a large-scale literature review, the collection and analysis of expert judgments of the validity of tests and inventories identified in the literature review, and the construction and administration of a Preliminary Battery of "off-the-shelf" tests and inventories. These activities served to direct the development of new predictor measures toward those abilities that seemed to hold most promise.

Three major types of new measures were developed: paper-and-pencil tests of cognitive ability (primarily in the spatial ability domain), paper-and-pencil inventories measuring temperament, biographical data, and vocational interest variables, and a set of computer-administered measures of perceptual/psychomotor abilities.

These new measures were developed in an iterative manner. The measures were subjected to three pilot tests with revisions occurring between each pilot test. All the measures were then collectively administered in a field test and final revisions were made.

During the pilot tests and the field test, several analyses and evaluations of the new measures were made. Score distributions and various types of test reliability were computed. The extent to which each new test or scale measured an ability not presently measured by the ASVAB (called uniqueness) was determined. The way in which the new measures related to each other and to the ASVAB subtests was analyzed. Investigations were made of the effect

of practice and idiosyncrasies of testing stations on computer-administered tests. The effects of faking on the temperament, biodata, and vocational interest measures were also investigated.

Findings:

The intended objectives of the research were realized. The newly developed predictor measures were shown to have adequate to excellent psychometric properties (that is, sufficiently large score distributions and acceptably high reliabilities), to be relatively unique (that is, to measure abilities not measured by the ASVAB), to be not unduly affected by practice, and not largely affected by faking in an applicant-like setting. Also, preliminary methods for detecting and correcting for faking were shown to be effective.

The final set of measures, called the Trial Battery, contains six paper-and-pencil, cognitive ability tests, 10 computer-administered tests of perceptual/psychomotor ability, and two paper-and-pencil inventories containing over 30 scales that measure temperament, biographical data, and vocational interests. The entire battery requires about 4 hours of time to administer.

Utilization of Findings:

The Trial Battery will be used in the Concurrent Validation Phase of Project A. Soldiers' scores on the Trial Battery will be compared to their scores on job performance criterion measures (also developed by Project A) to evaluate the validity of the Trial Battery and to evaluate the extent to which it improves the prediction of job performance over that achieved by the ASVAB.

CONTENTS

Page

OVERVIEW OF PROJECT A 1

CHAPTER 1. THEORETICAL APPROACH, RESEARCH DESIGN AND ORGANIZATION, AND DESCRIPTION OF INITIAL RESEARCH ACTIVITIES 1-1

TASK 2: APPROACH AND RESEARCH DESIGN 1-1
Theoretical Approach 1-1
Research Objectives 1-2
Research Design 1-4
Organization 1-6

TASK 2: PROGRESS SUMMARY 1-7

LITERATURE REVIEW 1-9
Purpose 1-9
Search Procedures 1-9
Literature Search Results 1-10
Screening of Predictors 1-10

EXPERT JUDGMENTS 1-13
Approach and Rationale 1-13
Identification of Predictor Variables 1-14
Identification of Criterion Variables 1-14
Subjects 1-16
Instructions and Procedures 1-16
Results 1-17

PRELIMINARY BATTERY 1-21
Purpose 1-21
Selection of Preliminary Battery Measures 1-21
Sample and Administration of Battery 1-23
Analyses 1-23

COMPUTER BATTERY DEVELOPMENT 1-25
Background 1-25
Phase 1. Information Gathering 1-25
Phase 2. Demonstration Battery 1-27
Phase 3. Selection/Purchase of Microprocessors and Development/Tryout of Software 1-28
Phase 4. Continued Software Development and Design/ Construction of a Response Pedestal 1-31

Page

PILOT TRIAL BATTERY 1-34
Identification of Measures 1-34
Pilot Tests and Field Tests 1-34

Chapter 1 References 1-36

CHAPTER 2. TEST DATA COLLECTION: PROCEDURES AND SAMPLES 2-1

PILOT TESTS 2-1
Pilot Test #1: Fort Carson 2-1
Pilot Test #2: Fort Campbell 2-3
Pilot Test #3: Fort Lewis 2-4
Summary of Pilot Tests 2-7

FIELD TESTS 2-11
Field Test of Pilot Trial Battery: Fort Knox 2-11
Summary 2-17

CHAPTER 3. COGNITIVE PAPER-AND-PENCIL MEASURES: PILOT TESTING 3-1

GENERAL 3-1
Target Population 3-1
Power vs. Speed 3-2
Reliability 3-3
Predictor Categories 3-4

SPATIAL VISUALIZATION 3-5
Spatial Visualization - Rotation 3-5
Spatial Visualization - Scanning 3-14

FIELD INDEPENDENCE 3-24

SPATIAL ORIENTATION 3-29

INDUCTION - FIGURAL REASONING 3-44

OVERALL ANALYSIS OF PILOT TEST RESULTS FOR COGNITIVE PAPER-AND-PENCIL MEASURES 3-54
Test Intercorrelations and Factor Analysis Results 3-54
Subgroup Analyses Results 3-59
Other Cognitive Tests 3-59

Chapter 3 References 3-63

Page

CHAPTER 4. COGNITIVE PAPER-AND-PENCIL MEASURES: FIELD TEST 4-1

ANALYSES OF DATA FROM FIELD TEST ADMINISTRATION 4-2
Mean Scores and Reliability Estimates 4-2
Gain Score Analysis 4-2
Covariance with ASVAB Subtests 4-8
Uniqueness Estimates of Cognitive Tests 4-11
Summary of Analyses 4-13

CHAPTER 5. PERCEPTUAL/PSYCHOMOTOR COMPUTER-ADMINISTERED MEASURES: PILOT TESTING 5-1

GENERAL 5-1
Test Development 5-1
Development of Response Pedestal 5-2

REACTION TIME (PROCESSING EFFICIENCY) 5-6
Simple Reaction Time: Reaction Time Test 1 5-6
Choice Reaction Time: Reaction Time Test 2 5-13

SHORT-TERM MEMORY 5-17
Memory Search Test 5-17

PERCEPTUAL SPEED AND ACCURACY 5-23
Perceptual Speed and Accuracy Test 5-23
Target Identification Test 5-32

PSYCHOMOTOR PRECISION 5-39
Target Tracking Test 1 5-39
Target Shoot Test 5-43

MULTILIMB COORDINATION 5-47
Target Tracking Test 2 5-47

NUMBER OPERATIONS 5-50
Number Memory Test 5-50

MOVEMENT JUDGMENT 5-53
Cannon Shoot Test 5-53

SUMMARY 5-56
Investigation of Machine Effects 5-56
Pilot Test Results: Comments 5-56

Chapter 5 References 5-60

Page

CHAPTER 6: PERCEPTUAL/PSYCHOMOTOR COMPUTER-ADMINISTERED MEASURES: FIELD TEST 6-1

PERCEPTUAL/PSYCHOMOTOR COMPUTERIZED TESTS ADMINISTERED 6-1

ANALYSIS OF DATA FROM FIELD TEST ADMINISTRATION 6-10
Field Test Scoring Procedures 6-10
Mean Scores and Reliability Estimates 6-13
Uniqueness Estimates of Computer-Administered Test Scores 6-14
Correlations with Video Game-Playing Experience 6-14
Effects of Differences in "Machine" or Computer Testing Station 6-14

EFFECTS OF PRACTICE ON SELECTED COMPUTER-ADMINISTERED TEST SCORES 6-19

COVARIANCE ANALYSES WITH ASVAB SUBTESTS AND COGNITIVE PAPER-AND-PENCIL TESTS 6-25

FACTOR ANALYSIS OF PTB COGNITIVE PAPER-AND-PENCIL MEASURES, PTB PERCEPTUAL-PSYCHOMOTOR COMPUTER-ADMINISTERED TESTS, AND ASVAB SUBTESTS 6-29

Chapter 6 References 6-32

CHAPTER 7. NON-COGNITIVE MEASURES: PILOT TESTING 7-1

GENERAL 7-1
Desired Characteristics 7-1
ABLE and AVOICE 7-2

TEMPERAMENT/BIODATA CONSTRUCTS 7-3
Adjustment 7-4
Dependability 7-5
Achievement 7-6
Physical Condition 7-6
Leadership (Potency) 7-7
Locus of Control 7-7
Agreeableness/Likeability 7-8
Response Validity Scales 7-8

ABLE REVISIONS BASED ON PILOT TESTING 7-11

PILOT TEST DATA FOR THE ABLE 7-13
Fort Campbell 7-13
Fort Lewis 7-17

Page

INTERESTS CONSTRUCTS 7-25
Realistic Interests 7-25
Conventional Interests 7-29
Social Interests 7-30
Investigative Interests 7-30
Enterprising Interests 7-31
Artistic Interests 7-31
Organizational Climate/Environment Scales 7-32
Expressed Interests Scale 7-32

AVOICE REVISIONS BASED ON PILOT TESTING 7-34

PILOT TEST DATA FOR THE AVOICE 7-35
Fort Campbell 7-35
Fort Lewis 7-35

SUMMARY 7-47

Chapter 7 References 7-48

CHAPTER 8. NON-COGNITIVE MEASURES: FIELD TESTS 8-1

ANALYSIS OF DATA FROM FIELD TEST ADMINISTRATION 8-4
Results of Data Quality Screening 8-4
Mean Scores and Reliability Estimates 8-4
Uniqueness Estimates for Non-Cognitive Measures 8-9
Factor Analysis of ABLE and AVOICE Scales 8-9

FAKABILITY INVESTIGATIONS 8-17
Purposes of the Faking Study 8-17
Procedure and Design 8-17
Faking Study Results - Temperament Inventory 8-20
Faking in An Applicant Setting 8-25
Faking Study Results - Interests Inventory 8-28

CONCLUDING COMMENTS 8-38

Chapter 8 References 8-39

CHAPTER 9. FORMULATION OF THE TRIAL BATTERY 9-1

REVISIONS TO THE PILOT TRIAL BATTERY 9-2
Changes to Cognitive Paper-and-Pencil Tests 9-4
Changes to Perceptual/Psychomotor Computer-Administered Tests 9-8
Changes to Non-Cognitive Measures (ABLE and AVOICE) 9-13

Page

DESCRIPTION OF THE TRIAL BATTERY AND SUMMARY CONTENTS 9-17

LIST OF APPENDIXES*

Page

APPENDIX A. DATA BASES SEARCHED A-1

B. COPIES OF ARTICLE REVIEW AND PREDICTOR REVIEW FORMS B-1

C. NAMES AND DEFINITIONS OF PREDICTOR AND CRITERION VARIABLES USED IN EXPERT JUDGMENT TASK C-1

D. SCALE NAMES AND NUMBER OF ITEMS IN EACH SCALE FOR THE PRELIMINARY BATTERY D-1

E. COMPUTERIZED MEASURES OBSERVED DURING SITE VISITS FOR ARI PROJECT A, SPRING 1983 . . . . E-1

LIST OF TABLES

Page

Table 2.1. Pilot Tests administered at Fort Carson, 17 April 1984 . . 2-2

2.2. Description of Fort Carson sample 2-3

2.3. Pilot Tests administered at Fort Campbell, 16 May 1984 . . 2-5

2.4. Description of Fort Campbell sample 2-6

2.5. Daily schedule for Fort Lewis pilot testing 2-7

2.6. Pilot Tests administered at Fort Lewis, 11-15 June 1984 . . 2-8

2.7. Description of Fort Lewis sample 2-9

---

*The Appendixes (F-H) that contain tests included in the Pilot Trial Battery or the Trial Battery are contained in a separate limited-distribution report: ARI Research Note 87-24, Test Appendixes to ARI Technical Report 739: Development and Field Test of the Trial Battery for Project A, April 1987.

		Page
Table 2.8.	Summary of pilot testing sessions for Pilot Trial Battery	2-10
2.9.	Race and gender of Fort Knox field test sample of the Pilot Trial Battery	2-15
2.10.	Military Occupational Specialities of Fort Knox Field Test sample of the Pilot Trial Battery	2-16
3.1.	Pilot Test results from Fort Lewis: Assembling Objects Test	3-8
3.2.	Pilot Test results from Fort Lewis: Object Rotation Test	3-12
3.3.	Pilot Test results from Fort Lewis: Path Test	3-17
3.4.	Pilot Test results from Fort Lewis: Maze Test	3-21
3.5.	Pilot Test results from Fort Lewis: Shapes Test	3-27
3.6.	Pilot Test results from Fort Lewis: Orientation Test 1	3-33
3.7.	Pilot Test results from Fort Lewis: Orientation Test 2	3-38
3.8.	Pilot Test results from Fort Lewis: Orientation Test 3	3-42
3.9.	Pilot Test results from Fort Lewis: Reasoning Test 1	3-47
3.10.	Pilot Test results from Fort Lewis: Reasoning Test 2	3-51
3.11.	Cognitive paper-and-pencil measures: Summary of Fort Lewis Pilot Test results	3-55
3.12.	Intercorrelations among the Ten Cognitive Paper-and-Pencil Measures	3-56
3.13.	Rotated Orthogonal Factor Solution for four factors	3-58
3.14.	Subgroup analyses of Cognitive Paper-and-Pencil Tests: White-black mean score differences in Pilot Test	3-60
3.15.	Subgroup analyses of Cognitive Paper-and-Pencil Tests: Male-female mean score differences in Pilot Test	3-61

		Page
Table 4.1.	Means, standard deviations, and reliability estimates for the Fort Knox Field Test of the Ten Cognitive Paper-and-Pencil Tests	4-7
4.2.	Gains on Pilot Trial Battery Cognitive Tests for persons taking tests at both Time 1 and Time 2	4-9
4.3.	Intercorrelations among the ASVAB subtests and the Cognitive Paper-and-Pencil Measures in the Pilot Trial Battery: For Knox sample	4-10
4.4.	Uniqueness estimates for Cognitive Tests in the Pilot Trial Battery (PTB) against tests in PTB and against tests in ASVAB	4-12
5.1.	Pilot Test results from Fort Lewis: Reaction Time Test 1 (simple reaction time)	5-8
5.2.	Mean correlations among Decision, Movement, and Total Times: Reaction Time Test 1	5-9
5.3.	Intercorrelations of dependent measures developed from Computer-Administered Tests: Fort Lewis Pilot Test	5-11
5.4.	Intercorrelations of Cognitive Paper-and-Pencil Tests and Computer-Administered Tests: Fort Lewis Pilot Test	5-12
5.5.	Pilot Test results from Fort Lewis: Reaction Time Test 2 (choice reaction time)	5-15
5.6.	Pilot Test results from Fort Lewis: Memory Search	5-19
5.7.	Pilot Test results from Fort Lewis: Overall Characteristics of Perceptual Speed and Accuracy Test	5-25
5.8.	Pilot Test results from Fort Lewis: Dependent measure scores from Perceptual Speed and Accuracy Test	5-29
5.9.	Intercorrelations among Perceptual Speed and Accuracy Test scores	5-30
5.10.	Pilot Test results from Fort Lewis: Target Identification Test	5-36
5.11.	Pilot Test results from Fort Lewis: Target Tracking Test 1	5-41

Page

Table 5.12. Pilot Test results from Fort Lewis: Target Shoot Test . . 5-46

5.13. Pilot Test results from Fort Lewis: Target Tracking Test 2 5-48

5.14. Means, standard deviations, and split-half reliability coefficients for 24 computer measure scores based on Fort Lewis Pilot Test data 5-57

5.15. Results of analyses of variance for machine effects: White and non-white males, Fort Lewis sample 5-58

6.1. Characteristics of the 19 dependent measures for Computer-Administered Tests: Fort Knox Field tests . . . 6-11

6.2. Uniqueness estimates for the 19 scores on Computer-Administered Tests in the Pilot Trial Battery against other computer scores and against ASVAB 6-15

6.3. Correlations between Computer Test scores and previous experience with video games 6-16

6.4. Effects of machine differences on Computer Test scores: Fort Knox Field Test 6-18

6.5. Effects of practice on selected Computer Test scores . . . 6-21

6.6. Gain scores and reliabilities for retest and practice groups 6-22

6.7. Intercorrelations among items 1-9, items 10-18, and items 19-27 of Target Tracking Tests 1 and 2 6-23

6.8. Intercorrelations among the ASVAB subtests and the Pilot Trial Battery Cognitive Paper-and-Pencil and Perceptual/Psychomotor Computer-Administered Tests: Fort Knox sample 6-26

6.9. Mean correlations, standard deviations, and minimum and maximum correlations between scores on ASVAB subtests and Pilot Trial Battery Tests of Cognitive, Perceptual, and Psychomotor abilities 6-28

6.10. Principal components factor analysis of scores of the ASVAB subtests, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer-Administered Tests 6-30

Page

Table 7.1. Temperament/Biodata Scales (by construct) developed for Pilot Trial Battery: ABLE - Assessment of Background and Life Experiences 7-4

7.2. Fort Campbell Pilot Test: ABLE Scale statistics 7-14

7.3. Fort Campbell Pilot Test: ABLE Scale intercorrelations . . 7-15

7.4. Fort Campbell Pilot Test: Varimax rotated principal factor analyses of 10 ABLE Scales 7-16

7.5. Fort Campbell Pilot Test: Correlations between ABLE constructs and scales and Personal Opinion Inventory (POI) marker variables 7-17

7.6. Fort Lewis Pilot Test: ABLE Scale statistics for total group 7-18

7.7. Fort Lewis Pilot Test: ABLE Scale means and standard deviations separately for males and females 7-19

7.8. Fort Lewis Pilot Test: ABLE Scale means and standard deviations separately for blacks and whites 7-20

7.9. Fort Lewis Pilot Test: ABLE Scale intercorrelations . . . 7-21

7.10. Fort Lewis Pilot Test: ABLE Scale intercorrelations with social desirability variance partialed out 7-23

7.11. Fort Lewis Pilot Test: Varimax rotated principal factor analyses of 10 ABLE Scales 7-24

7.12. Holland Basic Interest Constructs, and Army Vocational Interest Career Examination Scales developed for Pilot Trial Battery: AVOICE - Army Vocational Interest Career Examination 7-26

7.13. Additional AVOICE measures: Organizational Climate/ Environment and Expressed Interests Scales 7-27

7.14. Fort Campbell Pilot Test: AVOICE Scale statistics 7-36

7.15. Fort Lewis Pilot Test: AVOICE Scale statistics for total group 7-38

7.16. Fort Lewis Pilot Test: AVOICE means and standard deviations separately for males and females 7-40

Page

Table 7.17. Fort Lewis Pilot Test: AVOICE means and standard deviations separately for blacks and whites 7-42

7.18. Fort Lewis Pilot Test: AVOICE Scale intercorrelations . . 7-44

8.1. Fort Knox Field Test: Data quality screen results 8-5

8.2. Fort Knox Field Test: ABLE Scale score characteristics . . 8-6

8.3. Fort Knox Field Test: ABLE test-retest results 8-7

8.4. Fort Knox Field Test: AVOICE Scale score characteristics . 8-8

8.5. Uniqueness estimates for 11 ABLE Scales in the Pilot Trial Battery against other ABLE scores and against ASVAB 8-10

8.6. Uniqueness estimates for 24 AVOICE Scales in the Pilot Trial Battery against ASVAB 8-11

8.7. Summary of overlap of non-cognitive measures with other Pilot Trial Battery measures 8-12

8.8. Fort Knox Field Test: ABLE factor analysis 8-13

8.9. Fort Knox Field Test: AVOICE factor analysis 8-15

8.10. Faking Experiment, ABLE and AVOICE: Fort Bragg 8-19

8.11. Fakability Study, MANOVA results for ABLE Scales: Fort Bragg 8-21

8.12. Honesty and faking effects, ABLE Content Scales: Fort Bragg 8-22

8.13. Honesty and faking effects, ABLE Response Validity Scales: Fort Bragg 8-23

8.14. Effects of regressing out two Response Validity Scales (Social Desirability and Poor Impression) on faking condition, ABLE Content Scale Scores: Fort Bragg 8-24

8.15. Comparison of results from Fort Bragg honest, Fort Knox, and MEPS (recruits) ABLE Scales 8-27

Page

Table 8.16. Fakability Study, MANOVA results for AVOICE Combat-Related Scales: Fort Bragg 8-29

8.17. Fakability Study, MANOVA Results for AVOICE Combat-Support Scales: Fort Bragg 8-30

8.18. Effects of faking, AVOICE Combat Scales: Fort Bragg . . . 8-31

8.19. Effects of faking, AVOICE Combat Support Scales: Fort Bragg 8-32

8.20. Effects of regressing out Response Validity Scales (Unlikely Virtues and Poor Impression) on faking condition, AVOICE Combat Scales scores: Fort Bragg . . 8-33

8.21. Comparison of Fort Bragg honest, Fort Knox, and MEPS (recruits) AVOICE Combat-Related Scales 8-35

8.22. Comparison of Fort Bragg honest, Fort Knox, and MEPS (recruits) AVOICE Noncombat-Related Scales 8-36

9.1. Summary of changes to Cognitive Paper-and-Pencil Measures in the Pilot Trial Battery 9-5

9.2. Summary of changes to Computer-Administered Pilot Trial Battery Measures 9-6

9.3. Summary of changes to Pilot Trial Battery versions of Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE) 9-7

9.4. Summary of item reduction changes for ABLE and AVOICE . . 9-14

9.5. Number of items in Pilot Trial Battery and Trial Battery versions of ABLE Scales 9-15

9.6. Number of items in Pilot Trial Battery and Trial Battery versions of AVOICE Scales 9-16

LIST OF FIGURES

Page

Figure 1.1. Illustrative construct-oriented model 1-3

1.2. Flow chart of predictor measure development activities of Project A 1-5

1.3. Factors used to evaluate predictor measures for the Preliminary Battery 1-12

1.4. Hierarchical map of predictor space 1-18

1.5. Predictor categories discussed at IPR March 1984, linked to Pilot Trial Battery test names 1-35

2.1. Daily testing schedule for Fort Knox Field Test, weeks 1 and 2 2-12

2.2. Daily location schedule for Fort Knox Field Test, weeks 1 and 2 2-13

2.3. Daily schedule for Fort Knox Field Test, weeks 3 and 4 . . 2-14

3.1. Sample items from Assembling Objects Test 3-6

3.2. Distribution of items difficulty levels: Assembling Objects Test 3-9

3.3. Sample test items from Object Rotations Test 3-11

3.4. Distribution of item difficulty levels: Object Rotation Test 3-13

3.5. Sample items from Path Test 3-15

3.6. Distribution of item difficulty levels: Path Test 3-18

3.7. Sample items for the Maze Test 3-20

3.8. Distribution of item difficulty levels: Maze Test 3-22

3.9. Sample items from the Shapes Test 3-25

3.10. Distribution of item difficulty levels: Shapes Test . . . 3-28

3.11. Sample items from Orientation Test 1 3-31

Page

Figure 3.12. Distribution of item difficulty levels: Orientation Test 1 3-34

3.13. Sample items from Orientation Test 2 3-36

3.14. Distribution of item difficulty levels: Orientation Test 2 3-39

3.15. Sample items from Orientation Test 3 3-41

3.16. Distribution of item difficulty levels: Orientation Test 3 3-43

3.17. Sample items from Reasoning Test 1 3-46

3.18. Distribution of item difficulty levels: Reasoning Test 1 3-48

3.19. Sample items from Reasoning Test 2 3-50

3.20. Distribution of item difficulty levels: Reasoning Test 2 3-52

4.1. Description of Cognitive Paper-and-Pencil Measures in Field Test 4-3

5.1. Response pedestal for computerized tests 5-3

5.2. Key to flow diagrams of computer-administered tests . . . 5-5

5.3. Reaction Time Test 1 5-6

5.4. Reaction Time Test 2 5-14

5.5. Memory Test 5-18

5.6. Perceptual Speed and Accuracy Test 5-24

5.7. Type x Digit analysis of variance on Total Reaction Time 5-27

5.8. Type x Digit analysis of variance on Movement Time 5-28

5.9. Graphic displays of example items from the computer-administered Target Identification Test 5-33

Page

Figure 5.10. Target Identification Test 5-34

5.11. Distribution of 48 items on the revised Target Identification Test according to five parameters 5-38

5.12. Target Tracking Test 1 5-40

5.13. Target Shoot Test 5-44

5.14. Target Tracking Test 2 5-48

5.15. Number Memory Test 5-52

5.16. Cannon Shoot Test 5-54

6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test 6-2

6.2. Experimental design of the practice effects investigation 6-20

6.3. Items in the Computer Practice Battery used at the Fort Knox Field Test 6-20

7.1. Organizational climate/environment preference constructs, scales within constructs, and an item from each scale 7-33

8.1. ABLE scales organized by construct 8-2

8.2. AVOICE scales organized by construct 8-3

8.3. Debriefing Form used in the faking study at the Military Entrance Processing Station (MEPS) 8-26

8.4. Form filled out by MEPS recruits before debriefing 8-26

9.1. Guidelines for evaluating and retaining Pilot Trial Battery measures in order to produce the Trial Battery 9-3

9.2. Description of Trial Battery measures 9-18

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program which the U.S. Army has undertaken to develop an improved personnel selection and classification system for enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.
- Develop and validate new selection and classification measures.
- Validate intermediate criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), so that better informed reassignment and promotion decisions can be made throughout a soldier's career.
- Determine the relative utility to the Army of different performance levels across MOS.
- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The research design for the project incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from Army accessions in fiscal years (FY) 1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their subsequent performance in training and their scores on the first-tour Skills and Qualification Tests (SQT).

In the second iteration, a concurrent validation design will be executed with FY83/84 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS). The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered perceptual and psychomotor measures, will be administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

In the third iteration (the Longitudinal Validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

For both the concurrent and longitudinal validations, the sample of MOS was specially selected as a representative sample of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45 percent of Army accessions. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

Activities and progress during the first two years of the project were reported for FY83 in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37, and for FY84 in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 85-14. Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY85 is under preparation.

For administrative purposes, Project A is divided into five research tasks:

Task 1 -- Validity Analyses and Data Base Management
Task 2 -- Developing Predictors of Job Performance
Task 3 -- Developing Measures of School/Training Success
Task 4 -- Developing Measures of Army-Wide Performance
Task 5 -- Developing MOS-Specific Performance Measures

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries that will be used in the comprehensive Concurrent Validation program which is being initiated in FY85.

The present report is one of five reports prepared under Tasks 2-5 to report the development of the measures and the results of the field tests, and to describe the measures to be used in Concurrent Validation. The five reports are:

Task 2 -- "Development and Field Test of the Trial Battery for Project A," Norman G. Peterson, Editor, ARI Technical Report 739, May 1987.

Task 3 -- "Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS," by Robert H. Davis, et al., ARI Technical Report in preparation.

Task 4 -- "Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program," by Elaine D. Pulakos, and Walter C. Borman, Editors, ARI Technical Report 716, October 1985.

Task 5 -- "Development and Field Test of Task-Based MOS-Specific Criterion Measures," Charlotte H. Campbell, et al., ARI Technical Report 717, October 1985.

-- "Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS," Jody L. Toquam, et al., ARI Technical Report in preparation.

CHAPTER 1

THEORETICAL APPROACH, RESEARCH DESIGN AND ORGANIZATION, AND DESCRIPTION OF INITIAL RESEARCH ACTIVITIES

Norman G. Peterson

TASK 2: APPROACH AND RESEARCH DESIGN

As described in the Overview, Project A is organized into five research tasks, and activities of Task 2 are the focus of this report. Task 2's specific objective is the development and validation of new (or improved) selection and classification measures.

At present, the U.S. Army has a large number of jobs (called Military Occupational Specialties or MOS) and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. As obvious as these facts are, they need to be stated because they are the overriding facts that have to be addressed by Task 2 research.

One implication of these facts is that a highly varied set of individual differences' variables must be put into use if there is to be a reasonable chance of improving the present level of accuracy of predicting training performance, job performance, and attrition/retention in a substantial proportion, if not all, of those jobs. Much less evident is the particular content of that set of individual differences variables, and the way the set should be developed and organized.

A second, and perhaps less obvious, implication is the notion that new predictor measures must be appropriate for selecting persons who do not have the training and experience to immediately begin performing their assigned jobs. This is true partly because of the vast numbers of job positions that need to be filled, partly because of the kinds of jobs found in the Army (infantry, artillery, etc.), and partly because of the population of persons that the Army draws from (young high-school graduates with little or no specialized training and job experience).

Theoretical Approach

These considerations led us to adopt a construct-oriented strategy of predictor development, but with a healthy leavening from the content-oriented strategy. Essentially, we endeavored to build up a model of predictor space by (1) identifying the major, relatively independent domains or types of individual differences' constructs that existed; (2) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria; and (3) further selecting those constructs that appeared to be the "best bets" for incrementing (over present predictors) the prediction of the set of criteria of concern (i.e., training/job performance and attrition/retention in Army jobs).

Ideally, the model would, we hoped, lead to the selection of a finite set of relatively independent predictor constructs that were also relatively independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would predict all or most of the criteria, yet possess enough heterogeneity to yield powerful, efficient classification of persons into different jobs.

The development of such a model also had the virtue that it could be at least partially "tested" at many points during the research effort, and not just at the end, when all the predictor and criterion data are in. For example, we could examine the covariance of newly developed measures with one another and with the present predictors, notably the Armed Services Vocational Aptitude Battery (ASVAB). If the new measures were not relatively independent of the ASVAB and measures from other domains as predicted by the model, then we could take steps to correct that. Also, by constructing such a visible model, we thought that modifications and improvements could be implemented much more straightforwardly.

Figure I.1 shows an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and the predictor space are depicted. As mentioned earlier, a great deal of the work of Project A is devoted to the development of criterion measures, and we, on the predictor side, have taken advantage of the information coming from those efforts as it has become available.

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, on the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It is imperative that the development of such models be done very carefully and conservatively, and subjected frequently to reality testing; we have kept this firmly in mind. However, the possession of such a model enables one to state fairly clearly why such and such a predictor is being researched, and to check quickly, at least rationally, whether the addition of a predictor is likely to improve prediction.

Finally, the model is depicted as a matrix with a hierarchical arrangement of both the rows and the columns. We have found it useful to employ this hierarchical notion, because it allows us to think in terms of appropriate levels of specificity for a particular problem as we do the research, or for future applications of measures. (See Peterson and Bownas, 1982, for further discussion of this type of model.)

Research Objectives

This theoretical approach led to the delineation of seven, more concrete objectives of our research. These were:

1. Identify measures of human abilities, attributes, or characteristics which are most likely to be effective in predicting, prior to entry into the Army, successful soldier performance in general and in classifying persons into MOS where they will be most successful, with special emphasis on attributes not tapped by current pre-enlistment measures.

2. Design and develop new measures or modify existing measures of these "best bet" predictors.

		CRITERIA							
		Training Performance			Job Task Performance		Attrition/ Retention		
PREDICTORS		Pass/ Fail	Test Grades	Attendance	Common Tasks	Specific Tasks	Finish Term	Reenlist	Early Discharge
	Verbal	M*	H	L	M	M	L	L	L
Cognitive	Numerical	M	H	.	.	.	.		
	Spatial								
	Precision								
Psychomotor	Coordination								
	Dexterity								
	Dependability								
Temperament	Dominance								
	Sociability								
	Realistic								
Interests	Artistic								
	Social	.	.	.	M	M	M	L	L

*Denotes expected strength of relationship, High, Medium, Low.

Figure 1.1. Illustrative construct-oriented model.

3. Develop materials and procedures for efficiently administering experimental predictor measures in the field.

4. Estimate and evaluate the reliability of the new pre-enlistment measures and their vulnerability to motivational set differences, faking, variances in administrative settings, and practice effects.

5. Determine the interrelationships (or covariance) between the new pre-enlistment measures and current pre-enlistment measures.

6. Determine the degree to which the validity of new pre-enlistment measures generalizes across MOS, that is, proves useful for predicting measures of successful soldier performance across quite different MOS and, conversely, the degree to which the measures are useful for classification or the differential prediction of success across MOS.

7. Determine the extent to which new pre-enlistment measures increase the accuracy of prediction of success and the accuracy of classification into MOS over and above the levels of accuracy reached by current pre-enlistment measures.

Research Design

To achieve these objectives, we have followed the design depicted in Figure 1.2. There are 15 subtasks in our actual research plan, each tied to one or more of the activities or products shown in Figure 1.2.

Several things, we feel, are noteworthy about the design. First, five test batteries are mentioned: Preliminary Battery, Demo Computer Battery, Pilot Trial Battery, Trial Battery, and Experimental Battery. These appear successively in time and allow us to modify and improve our predictors as we gather and analyze data on each successive battery or set of measures.

Second, a large-scale literature review and a quantified expert judgment process were utilized early in the project to take maximum advantage of earlier research and accumulated knowledge and expert opinion. The expert judgment process was used to develop an early model of both the predictor space and the criterion space, and relied heavily on the information gained from the literature review. By using the model that resulted from analyses of the experts' judgments of the relationships between predictor constructs and criterion dimensions, we were able to develop, carefully and efficiently, measures of the most promising predictor constructs.

Third, the design includes both predictive (for the Preliminary and Experimental Batteries) and concurrent (for the Trial Battery) validation modes of data collection, although that is not obvious from Figure 1.2. Thus, we are able to benefit from the advantage of both types of designs,--that is, early collection and analysis of empirical criterion-related validities in the case of the concurrent design, and less concern about range restriction and experiential effects in the predictive design.

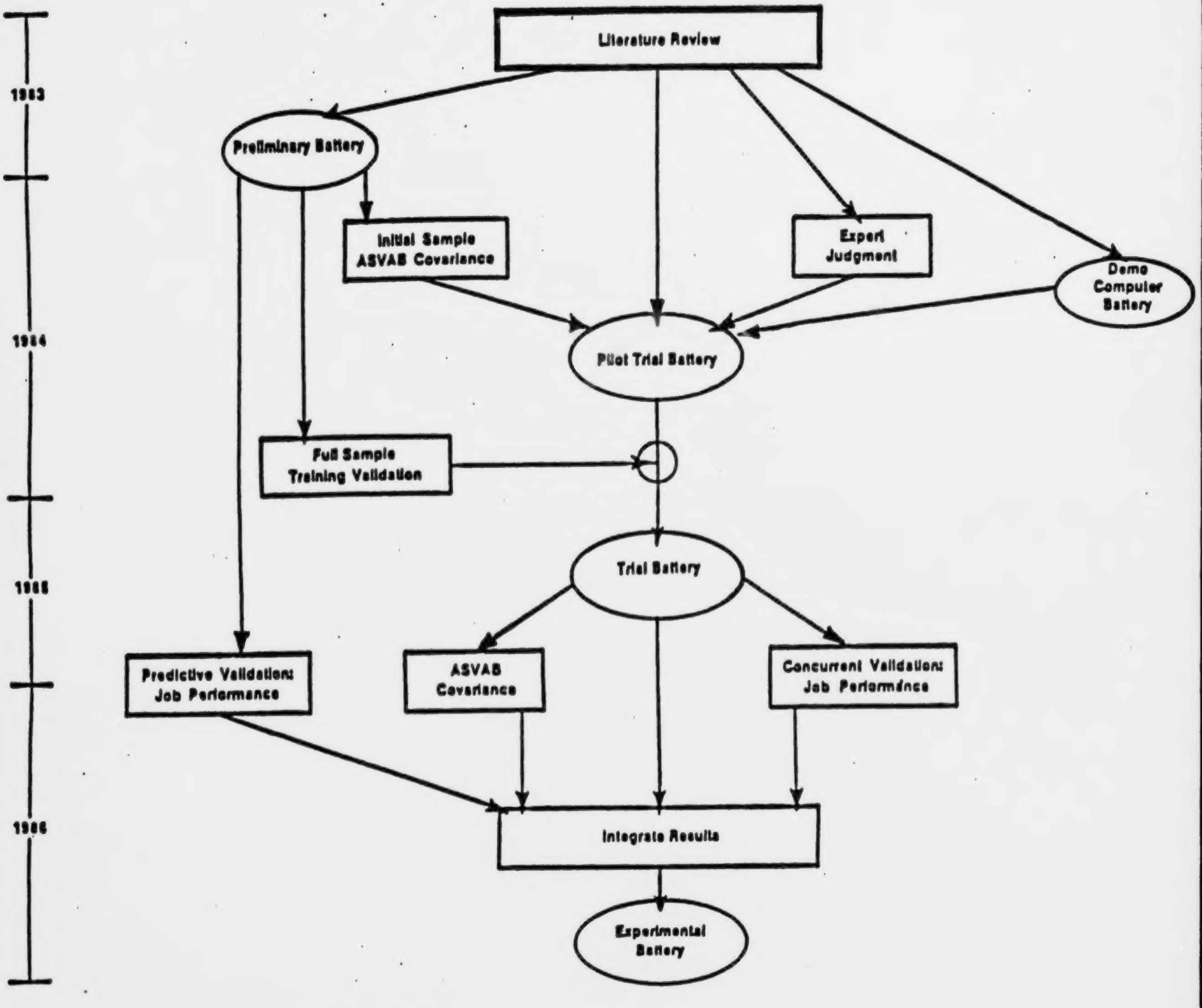

1983
1984
1985
1986
Literature Review
Preliminary Battery
Initial Sample
ASVAB Covariance
Expert
Judgment
Demo
Computer
Battery
Pilot Trial Battery
Full Sample
Training Validation
Trial Battery
Predictive Validation:
Job Performance
ASVAB
Covariance
Concurrent Validation:
Job Performance
Integrate Results
Experimental
Battery

Figure 1.2. Flow chart of predictor measure development activities of Project A.

Organization

We organized Task 2 researchers into three "domain teams" as we worked our way through this research design and toward the earlier described research objectives. One team concerned itself with the temperament, biographical data, and vocational interest variables and came to be called the "non-cognitive" team. Another team concerned itself with cognitive and perceptual kinds of variables and was called the "cognitive" team. The third team concerned itself with psychomotor and perceptual variables and was labeled the "psychomotor" team or sometimes the "computerized" team, since all the measures developed by that team were computer-administered.

TASK 2: PROGRESS SUMMARY

One gauge of progress is the degree to which the seven research objectives presented earlier have been accomplished. Following is a short summary of progress in terms of those objectives.

1. Identify "best bet" measures--This objective has been met. We sifted through a mountain of literature, translating the information onto a common form that enabled us to evaluate constructs and measures in terms of several psychomotor and pragmatic criteria. The results of that effort fed into the expert judgment process wherein 35 personnel psychologists provided the data necessary to develop our first model of the predictor space. After further review by experienced researchers in the Army and an advisory group, a set of "best bet" constructs was settled on. We also made some field visits to observe combat arms jobs first-hand,in addition to receiving criterion-side information from other Project A researchers; all of this information was very useful in developing new measures.

2. Develop measures of "best bet" predictors--This objective was accomplished by following the blueprint provided from the first objective. We carried out many small and not-so-small sample tryouts of these measures as they were developed, as is documented in the remainder of this report. The Trial Battery is the tangible product of meeting this objective.

3. Develop procedures for efficiently administering predictor measures--As anyone who has done research in military settings is aware, soldiers' time is precious and awarded research time is not to be squandered. We think we have developed and implemented effective methods for getting maximum quality and quantity of data out of our data collection efforts. The favorable results we have so far achieved in completeness and usefulness of data are due in large part, we think, to the attention paid to this objective.

4. Estimate reliability and vulnerability of measures--This objective has also been largely accomplished. We can report that analyses to date indicate that the new measures are psychometrically sound and acceptably invulnerable to the various sources of measurement problems--or we have devised some ways to adjust for such effects. However, more specifically targeted research would be useful in this area.

5. Determine the interrelationships between the new measures and current pre-enlistment measures--Work still remains on this objective, but the data collected to date show that the new measures have much variance that is not shared with the ASVAB, and that the across-domain shared variance is low (e.g., the new cognitive measures have low correlations with the non-cognitive measures).

6. and 7. Determine the level of prediction of soldier performance, classification efficiency, and incremental validity of the new measures--The jury is still out on these questions since the data that

will enable us to address these objectives have not yet been analyzed.

We turn now to a description of the initial research activities devoted to development of new predictors, specifically: literature review; expert judgments; development, administration, and analysis of the Preliminary Battery; and initial development of a computer battery. As Figure 1.2 shows, all of these activities led up to the development of the Pilot Trial Battery.

LITERATURE REVIEW

Purpose

The overriding purpose of the literature review was, simply put, to make maximum use of earlier research on the problem of accurately predicting job performance and classifying persons into jobs in such a way that both the person and the organization receive maximum benefits. More specifically, we wished to identify those variables or constructs, and their measures, that had proven effective for such purposes. As Figure 1.2 shows, the information obtained from the literature review was used in all the immediately succeeding research activities.

Search Procedures

The search was conducted by the three research teams, each responsible for a fairly broadly defined area of human abilities or characteristics: cognitive abilities; non-cognitive characteristics such as vocational interests, biographical data, and measures of temperament; and psychomotor/-physical abilities. While these domains were convenient for purposes of organizing and conducting literature search activities, they were not used as (nor intended to be) a final taxonomy of possible predictor measures.

The literature search was conducted in late 1982 and early 1983. In each of the three areas, the teams carried out essentially the same steps:

1. Compile an exhaustive list of possibly relevant reports, articles, books, or other sources.
2. Review each source and determine its relevancy for the project by examining the title and abstract (or other brief review).
3. Obtain the sources identified as relevant in the second step.
4. For relevant materials, carry out a thorough review and transfer relevant information onto special review forms developed for the project.

In the first step, several activities were designed to insure as comprehensive a list as possible. Several computerized searches of relevant data bases were done; Appendix A names and describes the data bases searched. Across all three ability areas, more than 10,000 sources were identified via the computer search. (Of course, many of these sources were identified as relevant in more than one area, and were thus counted more than once.)

In addition to the computerized searches, we obtained reference lists from recognized experts in each area, emphasizing the most recent research in the field. We also obtained several annotated bibliographies from military research laboratories. Finally, we scanned the last several years' editions of research journals that are frequently used in each ability area, as well as more general sources such as textbooks, handbooks, and appropriate chapters in the *Annual Review of Psychology* (which reviews the most recent research in a number of conceptually distinct areas of psychology).

The vast majority of the sources identified were not relevant to our purpose--that is, the identification and development of promising measures for personnel selection in the U.S. Army. These nonrelevant sources were weeded out in Step 2. The relevant sources were obtained and reviewed, and two forms were completed for each source: an Article Review form and a Predictor Review form (several of the latter could be completed for each source). These forms were designed to capture, in a standard format, the essential information about the reviewed sources, which varied considerably in their organization and reporting styles.

The Article Review form contained seen sections: citation, abstract, list of predictors (keyed to the Predictor Review forms), description of criterion measures, description of sample(s), description of methodology, other results, and reviewer's comments. The Predictor Review form also contained seven sections: description of predictor, reliability, norms/descriptive statistics, correlations with other predictors, correlations with criteria, adverse impact/differential validity/test fairness, and reviewer's recommendations (about the usefulness of the predictor). Each predictor was tentatively classified into an initial working taxonomy of predictor constructs (based primarily on the taxonomy described in Peterson and Bownas, 1982). Appendix B contains copies of these two forms.

Literature Search Results

The Review forms and the actual sources that had been located were used in two primary ways. First, three working documents were written, one for each of the three areas. (These work documents were put into ARI Research Note form: Toquam, Corpe, Dunnette and Keyes, in preparation; McHenry and Rose, in preparation; Hough, Kampe, and Barge, in preparation.) These documents identified and summarized the literature with regard to issues important to the research being conducted, the most appropriate organization or taxonomy of the constructs in each area, and the validities of the various measures for different types of job performance criteria. Second, the predictors identified in the review were subjected to further, structured scrutiny in order to (1) select tests, and inventories to make up the Preliminary Battery, and (2) select the "best bet" predictor constructs to be used in the expert judgment research activity.

Screening of Predictors

An initial list was compiled of all predictor measures that seemed even remotely appropriate for Army selection and classification. This list was further screened by eliminating measures according to several "knockout factors: (1) measures developed for a single research project only; (2) measures designed for a narrowly specified population/occupational group (e.g., pharmacy students); (3) measures targeted toward younger age groups; (4) measures requiring special apparatus for administration; (5) measures requiring unusually long testing times; (6) measures requiring difficult or subjective scoring; and (7) measures requiring individual administration.

Knockout factor (4) was applicable only with regard to screening for the Preliminary Battery, which could not have any computerized tests or other apparatus since it was to be administered early in the project, before such testing devices could be developed. Factor (4) was not applied with regard to screening measures for inclusion in the expert judgment process.

Application of knockout factors resulted in a second list of candidate measures. Each of these measures was evaluated on the 12 factors shown in Figure 1.3, by at least two researchers. (A 5-point rating scale was applied to each of the 12 factors.) Discrepancies in ratings were resolved by discussion. We point out that there was not always sufficient information for a variable to allow a rating on all factors.

This second list of measures, each with a set of evaluations, was input to (1) the final selection of measures for the Preliminary Battery and (2) the final selection of constructs to be included in the expert judgment process, to which we now turn.

1. Discriminability - extent to which the measure has sufficient score range and variance, i.e., does not suffer from ceiling and floor effects with respect to the applicant population.
2. Reliability - degree of reliability as measured by traditional psychometric methods such as test-retest, internal consistency, or parallel forms reliability.
3. Group Score Differences (Differential Impact) - extent to which there are mean and variance differences in scores across groups defined by age, sex, race, or ethnic groups; a high score indicates little or no mean differences across these groups.
4. Consistency/Robustness of Administration and Scoring - extent to which administration and scoring is standardized, ease of administration and scoring, consistency of administration and scoring across administrators and locations.
5. Generality - extent to which predictor measures a fairly general or broad ability or construct.
6. Criterion-Related Validity - the level of correlation of the predictor as a measure of job performance, training performance and turnover/attrition.
7. Construct Validity - the amount of evidence existing to support the predictor as a measure of a distinct construct (correlational studies, experimental studies, etc.).
8. Face Validity/Applicant Acceptance - extent to which the appearance and administration methods of the predictor enhance or detract from its plausibility or acceptability to laymen as an appropriate test for the Army.
9. Differential Validity - existence of significantly different criterion-related validity coefficients between groups of legal or societal concern (race, sex, age); a high score indicates little or no differences in validity for these groups.
10. Test Fairness - degree to which slopes, intercepts, and standard errors of estimate differ across groups of legal or societal concern (race, sex, age) when predictor scores are regressed on important criteria (job performance, turnover, training); a high score indicates fairness (little or no differences in slopes, intercepts, and standard errors of estimate).
11. Usefulness of Classification - extent to which the measure or predictor will be useful in classifying persons into different specialties.
12. Overall Usefulness for Predicting Army Criteria - extent to which predictor is likely to contribute to the overall or individual prediction of criteria important to the Army (e.g., AWOL, drug use, attrition, unsuitability, job performance, and training).

Figure 1.3. Factors used to evaluate predictor measures for the Preliminary Battery.

EXPERT JUDGMENTS

Approach and Rationale

The approach used in the expert judgment process was to (1) identify criterion categories, (2) identify an exhaustive range of psychological constructs that may be potentially valid predictors of those criterion categories, and (3) obtain expert judgments about the relationships between the two. Schmidt, Hunter, Croll, and McKenzie (1983) showed that pooled expert judgments, obtained from experienced personnel psychologists, were as accurate in estimating the validity of tests as actual, empirical criterion-related validity research using samples of hundreds of subjects. That is, experienced personnel psychologists are effective "validity generalizers" for cognitive tests. They do tend to underestimate slightly the true validity as obtained from empirical research.

Hence, one way to identify the "best best" set of predictor variables and measures is to use a formal judgment process employing experts such as that followed by Schmidt et al. (1983). Peterson and Bownas (1982) provide a complete description of the methodology, which has been used successfully by Bownas and Heckman (1976), Peterson, Houston, Bosshardt, and Dunnette (1977), Peterson and Houston (1980), and Peterson, Houston, and Rosse (1984) to identify predictors for the jobs of firefighter, correctional officer, and entry-level occupations (clerical and technical), respectively. Descriptive information about a set of predictors and the job performance criterion variables is given to "experts" in personnel selection and classification, typically personnel psychologists. These experts estimate the relationships between predictor and criterion variables by rating or directly estimating the value of the correlation coefficients.

The result is a matrix with predictor and criterion variables as the columns and rows, respectively. Cell entries are experts' estimates of the degree of relationship between the particular predictors and various criteria. The interrater reliability of the experts' estimates is checked first. If the estimate is sufficiently reliable (previous research shows values in the .80 to .90 range for about 10 to 12 experts), the matrix of predictor-criterion relationships can be analyzed and used in a variety of ways. By correlating the columns of the matrix, the covariances of the predictors can be estimated on the basis of the profiles of their estimated relationships with the criteria. These covariances can then be factor analyzed to identify predictors that function similarly in predicting performance criteria. Similarly, the criterion covariances can be examined to identify clusters of criteria predicted by a common set of predictors.

Such procedures help identify redundancies and overlap in the predictor set. The common sets or clusters of predictors and of criteria are an important product for several reasons. First, they provide an efficient means of summarizing the data generated by the experts. Second, the summary form allows easier comparison with the results of meta-analyses of criterion-related validity coefficients. Conflicting or absent evidence is a sure guide to important research questions. Certain clusters may have to be reconfigured because of new data. Third, less direct but potentially more important, these clusters provide a model or theory of predictor-criterion performance space. This model serves as an informative guide to development of a set of predictors that should be efficient and valid, at least insofar as the informed opinion of knowledgeable experts can propel one in that direction.

To carry out the expert judgment activity, we had to identify predictor and criterion variables and prepare materials that would enable the experts to provide reliable estimates of validity.

Identification of Predictor Variables

The list of predictor variables that had been evaluated on 12 relevant factors (see Literature Review, *Screening of Predictors*) was used to identify the predictors for the expert judgment process. Variables were included if they received generally high evaluations and if they added to the comprehensiveness of coverage for a particular domain of predictor variables. At this point, we began to depart somewhat from the initial predictor taxonomy used in the literature review, and to create a new one that we felt best represented the entire predictor domain relevant to our Army goal. There were 53 members in the final set of predictor variables. (The names and definitions of these variables are shown in Appendix C.)

Materials describing each of the 53 variables were prepared. The expert judges were experienced psychologists who were generally familiar with psychometric information and, in varying degrees, knowledgeable about the 53 variables in our final list. Therefore, the descriptive material was designed to transmit a large amount of information as concisely as possible.

Each packet contained a sheet that named and defined the variable, described how it was typically measured, and summarized the reliability and validity of the selected measures of the variable. Following this sheet were descriptions of one or more specific measures, including the name of the test, its publisher, the variable it was designed to measure, a description of the items and the number of items on the test (in most cases, sample items were included), a brief description of the administration and scoring of the test, and brief summaries of studies of the reliability and validity of the measure.

Identification of Criterion Variables

Several types of criterion variables were identified. They included a set of specific job task criterion categories, a set that described performance in initial Army Training, and a set of generalized Army effectiveness categories.

Specific Job Task Categories. Short of enumerating all job tasks in the nearly 240 entry-level job specialties, the nature of the performance domain had to be characterized in a way that was at once comprehensive, understandable, and usable by judges. Since many jobs share similar tasks, the abstraction of generic task categories was possible. Two approaches were tried; we report here only on the method chosen.

This approach was based on more general job descriptions of a representative sample of 111 jobs that had been previously clustered by personnel experts familiar with Army jobs. Twenty-three clusters had been identified. Criterion categories were developed by reviewing the descriptions of the jobs in these clusters to determine common job activities. Emphasis was placed on determining what a soldier in each job might be observed doing and what he or she might be trying to accomplish. The categories were constructed to connote a set of actions that typically occur together (e.g., transcribe, annotate, sort, index, file, retrieve) leading to some common objective (e.g., record and file information). Criterion categories often included reference to the use of equipment or other objects.

Once criterion categories were identified for the common actions in the 23 clusters, additional categories were identified to cover unique aspects of jobs in the sample of 111. In all, 53 categories were generated. Most of the categories applied to several jobs, and most of the jobs were characterized by activities from several categories. Their names and definitions are shown in Appendix C.

Performance in Initial Army Training. Two sources of information were used to identify appropriate training performance variables: archival records of soldiers' performance in training were examined, and trainers were interviewed. This information was obtained for eight MOS: Radio/Teletype Operator, MANPADS Crewman, Light Vehicle/Power Generation Mechanic, Motor Transport Operator, Food Service Specialist, M60 and M1 Armor Crew, Administrative Specialist, and Unit Supply Specialist. These specialties represented a heterogeneous group with respect to type of work and were, for the most part, high-density MOS.

The review of archival records was intended to identify the type of measures used to evaluate training performance, since the content was, obviously, specific to each MOS.

Five or six trainers were interviewed for each MOS, using a modified critical incidents approach. Trainers were asked, "What things do trainees do that tell you they are good (or bad) trainees?" Generally, trainers responded with fairly broad, trait-like answers and appropriate follow-up questions were used to obtain more specific, behaviorally oriented information.

After examining the archives and conducting the interviews, we pooled and categorized the information from both sources. We found much overlap across MOS in the way training performance was evaluated. Furthermore, we could not include content-specific variables since this would require several hundred training performance variables (one for each MOS, at least). Nor did we wish to do so, since the task or MOS-specific performance variance was covered elsewhere, as described above.

In the end, we decided that four variables adequately represented training performance. Their names and definitions are shown in Appendix C.

Generalized Army Effectiveness Categories. The identification of these variables was carried out in three steps. First, we developed a preliminary conceptual model based on relevant theory and empirical findings. Second, empirical research using the inductive behavioral analysis method was carried out to verify and modify the preliminary model. Finally, several criterion variables that are common across all MOS but are not behavioral in nature were added to the final list. We briefly summarize those steps here; a more complete description can be found in a paper by Borman, Motowidlo, and Hanser (1983).

The preliminary model revolved around three concepts: organizational commitment, organizational socialization, and morale. Each of these was thought to contribute to generalized Army effectiveness. Consideration of theory and research in these areas led to the identification and definition of 15 general Army effectiveness dimensions.

Behavioral analysis workshops were employed in order to verify and extend this model. Persons knowledgeable about a job were asked to generate behavioral examples of effective and ineffective performance in all aspects of the job. Army NCOs and officers generated several hundred examples, which were then content analyzed by Project A staff. The resulting categories were compared to the dimensions in the preliminary model. There was considerable overlap, but some modifications were made to the model dimensions. Nine general effectiveness behavioral dimensions were named and defined; these are shown in Appendix C.

In the final step, six more criterion variables indicating general effectiveness were added; they are also named and defined in Appendix C. The first two, " Survive in the field" and "Maintain physical fitness," were added because they are expected of all soldiers but did not emerge elsewhere. The last four are all important "outcome" criterion variables. That is, they represent outcomes of individual behavior that have negative or positive value to the Army, but the outcomes could occur because of a variety of individual behaviors.

In all, then, 72 criterion variables were identified and defined for use in the expert judgment task.

Subjects

The experts who served as judges were 35 industrial, measurement, or differential psychologists with experience and knowledge in personnel selection research and/or applications. Each expert was an employee of or consultant to one of the four organizations involved in Project A: U.S. Army Research Institute, Personnel Decisions Research Institute, Human Resources Research Organization, American Institutes for Research. Not all of the employees were directly involved with Project A although all of the consultants were.

Instructions and Procedures

Detailed instructions were provided for each judge along with the materials describing the predictor and criterion variables. Information was provided on the concept of "true validity," criterion-related validity corrected for such artifacts as range restriction and unreliability, and unaffected by variation in sample sizes. Judges were asked to estimate the level of true validity rather than estimated validity, on a 9-point scale. A rating of "1" meant a true validity in the range of .00 to .10; "2", .11 to .20; and so forth, to "9", .81 to .90.

Descriptions of the 53 predictor variables had been divided into three groups (A, B, and C, two groups of 18 and one of 17). The 72 criterion descriptions were in one group. The judges were encouraged to skim the materials for a few predictors and for all the criteria before beginning the rating task.

Each judge then estimated the validity of each predictor for each criterion. The order of the predictor groups (A, B, C) was counterbalanced across judges, with about one-third of the 35 judges beginning with Group A (Predictors 1-18), another one-third with Group B (Predictors 19-36), and the rest with Group C (Predictors 37-53).

Ratings were made on separate Judgment Record Sheets. Before making any judgments about a predictor, the expert was to read the description and review the examples given to measure it; judgments were to be made about the predictor as a construct, not about the variable as measured by any specific instrument. The judge was then to read the description of the first criterion and to estimate the validity of that predictor for that criterion. Judgments could be either positive or negative; positive signs were not to be entered. The judge was then to read the description of the second criterion and rate the validity of the same predictor for that criterion. The judge was to estimate the validities of the first predictor variable for all 72 criteria before moving to the next predictor.

All judges completed the task during the first week of October 1983.

Results

A number of analyses were carried out: reliability of the judgments, means and standard deviations of the estimated validities within each predictor/criterion cell and for various marginal values, and factor analyses of the predictors (based on their validity profiles across the criteria) and the criteria (based on their validity profiles across the predictors).

The estimated validities were highly reliable when averages were used. The reliability of the mean estimated cell validities was .96. The factor analyses were based on these cell means. The most pertinent analysis for purposes of this report concerns the factor analysis of the predictors.

Factor solutions with two through 24 factors were calculated. The nine-factor solution was selected as most meaningful. Eight of the nine factors were interpretable; one was not interpreted. The eight interpretable factors were named: Cognitive Abilities, Visualization/Spatial, Information Processing, Mechanical, Psychomotor, Social Skills, Vigor, Motivation/Stability.

These eight factors appeared to be composed of 21 clusters, based on the profile of loadings of each predictor variable across all the factors. This hierarchical structure of the predictor variables is shown in Figure 1.4. Inspection of the profiles clarifies the meanings both of the factors and of the clusters, as follows.

The eight predictor factors divide the predictor domain into reasonable-appearing parts. The first five refer to abilities and skills in the cognitive, perceptual, and psychomotor areas while the last three refer to traits or predispositions, in the noncognitive area. Most of the representative measures of the constructs defining the first five factors are of

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension	A. Verbal Ability/ General Intelligence	COGNITIVE ABILITIES
5. Reading Comprehension		
16. Ideational Fluency		
18. Analogical Reasoning		
21. Omnibus Intelligence/Aptitude		
22. Word Fluency		
4. Word Problems	B. Reasoning	
8. Inductive Reasoning: Concept Formation		
10. Deductive Logic		
2. Numerical Computation	C. Number Ability	
3. Use of Formula/Number Problems		
12. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory	J. Memory	
17. Follow Directions		
19. Figural Reasoning	F. Closure	
23. Verbal and Figural Closure		
6. Two-dimensional Mental Rotation	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
7. Three-dimensional Mental Rotation		
9. Spatial Visualization		
11. Field Dependence (Negative)		
15. Place Memory (Visual Memory)		
20. Spatial Scanning		
24. Processing Efficiency	G. Mental Information Processing	INFORMATION PROCESSING
25. Selective Attention		
26. Time Sharing		
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests	M. Realistic vs. Artistic Interests	
51. Artistic Interests (Negative)		
28. Control Precision	I. Steadiness/Precision	PSYCHOMOTOR
29. Rate Control		
32. Arm-hand Steadiness		
34. Aiming		
27. Multilimb Coordination	D. Coordination	
35. Speed of Arm Movement		
30. Manual Dexterity	K. Dexterity	
31. Finger Dexterity		
33. Wrist-Finger Speed		
39. Sociability	Q. Sociability	SOCIAL SKILLS
52. Social Interests		
50. Enterprising Interests	R. Enterprising Interests	
36. Involvement in Athletics and Physical Conditioning	T. Athletic Abilities/Energy	VIGOR
37. Energy Level		
41. Dominance	S. Dominance/Self-esteem	
42. Self-esteem		
40. Traditional Values	N. Traditional Values/Conventionality/Non-delinquency	MOTIVATION/ STABILITY
43. Conscientiousness		
46. Non-delinquency		
53. Conventional Interests		
44. Locus of Control	O. Work Orientation/Locus of Control	
47. Work Orientation		
38. Cooperativeness	P. Cooperation/Emotional Stability	
45. Emotional Stability		

Figure 1.4. Hierarchical map of predictor space.

maximal performance while most of the representative measures of the last three factors are of typical performance, with the exception of the interest variables.

The first four factors, which include 11 clusters of 29 predictor constructs or variables, are cognitive-perceptual in nature. The first factor, labeled "Cognitive Abilities," includes seven clusters, five of which appear to consist of more traditional mental test variables: Verbal Ability/General Intelligence, Reasoning, Number Ability, Memory, Closure. The Perceptual Speed and Accuracy cluster is linked to measures having a long history of inclusion in traditional mental tests. The seventh cluster, Investigative Interests, refers to no cognitive test at all but does tap interest in things intellectual, the abilities for which are evaluated in this factor.

The second factor, Visualization/Spatial, consists of only one cluster but includes six constructs which have some history of assessment of spatial ability. Two of the clusters from the Cognitive Abilities factor, Reasoning and Closure, have some affinity to this second factor, as may be seen in the factor analysis data. This may be due to the tasks used to illustrate the assessment of the constructs, which are to solve problems of a visual and nonverbal nature. The third factor, Information Processing, also consists of only one cluster, with the three constructs referring more directly to cognitive-perceptual functioning rather than accumulated knowledge and/or structure.

The fourth factor, Mechanical, includes two clusters, one of which consists only of the construct of Mechanical Comprehension while the other is, again, an interest cluster consisting of a positive loading for Realistic Interests and negative loading for Artistic Interests.

The fifth factor, Psychomotor, consists of three clusters which include the nine psychomotor constructs. The first cluster, Steadiness/Precision, refers to aiming and tracking tasks, where the target may move steadily or erratically. The second cluster, Coordination, indexes the large-scale complexity of the response required in a psychomotor task while the third factor, Dexterity, appears to index the small-scale complexity of responses.

The remaining three factors, noncognitive in character, refer more to interpersonal activities. The Social Skills factor consists of two clusters. The first, Sociability, refers to a general interest in people while the second, Enterprising Interests, refers to a more specific interest in working successfully with people. The seventh factor is called "Vigor" as it includes two clusters that both refer to general activity level. The first, Athletic Abilities/Energy, includes two constructs which point towards a physical perspective while the second cluster, Dominance/Self-Esteem, points toward a psychological perspective.

The eighth and last factor, Motivation/Stability, includes three clusters or facets. The first, Traditional Values, includes both temperament measures and interest scales, and refers to being rule-abiding and a good citizen. The second cluster, Work Orientation, refers to temperament measures which index attitudes towards the individual vis-a-vis his/her efforts in the world. The third cluster, Cooperation/Stability, appears to refer to skill in getting along with people, including getting along with oneself in a healthy manner.

The expert judgment task resulted in a hierarchical model of predictor space that served as a guide for the development of new, pre-enlistment measures (the Pilot Trial Battery, see Figure 1.2) for Army enlisted ranks. (Wing, Peterson, and Hoffman, 1984, provide a detailed presentation of the expert judgment process and results.) This model was not the only set of information that guided the development of the Pilot Trial Battery, however. We turn now to the other major source of guidance, the development, administration, and initial analyses of the Preliminary Battery.

PRELIMINARY BATTERY

Purpose

The Preliminary Battery (PB) was conceived of as a set of proven "off-the-shelf" measures of predictors that overlapped very little with the Army's current pre-enlistment predictors. There were two primary reasons for developing and administering a Preliminary Battery. First, the collection of data on a number of predictors that represent the types of predictors not currently in use by the Army would allow an early determination of the extent to which such predictors contributed unique variance, that is, measured attributes not measured by current pre-enlistment predictors. This information would be useful for guiding the development of new predictors into areas most likely to be useful for increasing the accuracy of prediction and classification.

Second, the collection of predictor data (from soldiers in training) early in the project allowed the conduct of a predictive validity investigation much earlier in the project than if we were to wait until the Trial Battery was developed (see Figure 1.2). Thus, the extent to which the different (from ASVAB) constructs represented in the Preliminary Battery added to the prediction of training success and effectiveness of job performance could be ascertained via a predictive design approximately 18 months and 36 months after Project A began, rather than many months later than that.

Selection of Preliminary Battery Measures

As described earlier, the literature review identified a large set of predictor measures, each with ratings by the researchers on 12 psychometric and substantive evaluation factors (see Figure 1.3). These ratings were used to select a smaller set of measures as serious candidates for inclusion in the Preliminary Battery. Two major practical constraints came into play: (1) no apparatus or individualized testing methods could be used because of the relatively short time available to prepare for battery administration, and the fact that the battery would be administered to a large number of soldiers (several thousand) over a 9-month period by relatively unsophisticated test administrators, and (2) only 4 hours were available for testing.

Task 2 researchers made an initial selection of "off-the-shelf" measures, but there were still too many measures for the time available. The tentative list was referred to the Army Research Institute scientists responsible for Task 2 specifically, and Project A generally, and to the Project A Director and Principal Investigator. The available information about each measure (construct measured, psychometric characteristics, type of job performance criteria it had predicted or was thought likely to predict) was presented and discussed. The set of measures selected was then reviewed by several consultants external to Project A, who had been retained for their expertise in various predictor domains. These experts made several "fine-tuning" suggestions.

The Preliminary Battery included the following:

- Eight perceptual-cognitive measures

  - Five from the Educational Testing Service (ETS) French Kit (Ekstrom, French, and Harman, 1976)
  - Two from the Employee Aptitude Survey (EAS) (Ruch and Ruch, 1980)
  - One from the Flanagan Industrial Tests (FIT) (Flanagan, 1965)

- Eighteen scales from the Air Force Vocational Interest Career Examination (VOICE) (Alley and Matthews, 1982)

- Five temperament scales adapted from published scales
  - Two from the Differential Personality Questionnaire (DPQ) (Tellegen, 1982)
  - One from the California Psychological Inventory (CPI) (Gough, 1975)
  - The Rotter I/E scale (Rotter, 1966)
  - Validity scales from both the DPQ and the Personality Research Form (PRF) (Jackson, 1967)

- Owen's Biographical Questionnaire (BQ) (Owens and Schoenfeldt, 1979). The BQ could be scored for either 11 scales for males or 14 for females, based on Owen's research, or for 18 predesignated, combined-sex scales developed for this research and called Rational Scales. The rational scales had no item on more than one scale; some of Owen's scales included items on more than one scale. Items tapping religious or socio-economic status were deleted from Owens' instrument for this use, and items tapping physical fitness and vocational-technical course work were added.

Appendix D shows all the scale names and numbers of items for the Preliminary Battery.

In addition to the Preliminary Battery, scores were available for the Armed Services Vocational Aptitude Battery, which all soldiers take prior to entry into service. ASVAB's ten subtests are named below, with the test acronym and number of items in parentheses:

> Word Knowledge (WK:35), Paragraph Comprehension (PC:15), Arithmetic Reasoning (AR:30), Numerical Operations (NO:50), General Science (GS:25), Mechanical Comprehension (MC:25), Math Knowledge (MK:25), Electronics Information (EI:20), Coding Speed (CS:84), Auto-Shop Information (AS:25).

All but NO and CS are considered to be power tests; the two exceptions are speeded. Prior research (in Kass, Mitchell, Grafton, & Wing, 1983) has shown the reliability of the subtests to be within expectable limits for cognitive tests of this length (i.e., .78-.92).

Sample and Administration of Battery

The Preliminary Battery was administered to soldiers entering Advanced Individual Training (AIT) for four MOS: 05C, Radio Teletype Operator (MOS code was later changed to 31C); 19 E/K, Armor Crewman; 63B, Vehicle and Generator Mechanic; and 71L, Administrative Specialist. Almost all soldiers entering AIT for these MOS during the period 1 October, 1983 to 30 June, 1984 completed the Preliminary Battery. We are here concerned only with the sample of soldiers who completed the battery from 1 October, 1983 to 1 December, 1983, approximately 2,200 soldiers.

The battery was administered at five training posts by civilian or military staff already employed on site. Task 2 staff traveled to these sites to deliver battery administration manuals and to train the persons who would administer the battery. A full day of training was provided, including a complete reading of the administration manual, role-playing practice in reading test and inventory instructions, completion of all tests and inventories by the administrators, and question-and-answer sessions about each chapter of the administration manual. Thereafter, Task 2 staff contacted each post each week by telephone to receive progress reports and answer questions. Administrators at posts also called Task 2 staff whenever they had questions. The experience in training battery administrators and monitoring the administration over the nine-month period provided useful information for the data collection efforts involving the Pilot Trial Battery and Trial Battery.

We note here that the Preliminary Battery was administered to a sample of 40 soldiers at Fort Leonard Wood prior to its implementation in order to test the instructions, timing, and other administration procedures. The results of this tryout were used to adjust the procedures, prepare the manual, and identify topics to be emphasized during administrator training.

Analyses

An initial set of analyses was performed on the Preliminary Battery data to inform the development of the Pilot Trial Battery (PTB). (The PTB was intended to include newly developed tests and inventories that would measure the important abilities and traits identified via the literature review and expert judgment process. These PTB measures would be piloted and field tested and then revised to become the Trial Battery. See Figure 1.2 for a flow chart showing the sequencing of the various batteries.) We summarize those findings here. They are more completely reported in Hough, Dunnette, Wing, Houston, and Peterson (1984).

Three types of analyses were done. First, the psychometric characteristics of each scale were explored to pinpoint possible problems with the measures or the construct being measured, so those problems could be avoided when the Pilot Trail Battery measures were developed. These analyses included descriptive statistics, item analyses (including numbers of items attempted in the time allowed), internal consistency reliability estimates, and, for the temperament inventory, percentage of subjects failing the scales intended to detect random or improbable response patterns.

Second, the covariances of the scales within and across the various conceptual domains (i.e., cognitive, temperament, biographical data, and vocational interest) were investigated to detect excessive redundancy among the PB measures, especially across the domains. If such redundancies were detected, then steps could be taken to avoid such a problem in the Pilot Trial Battery. Third, the covariances of the PB scales with ASVAB measures were studied to identify any PB constructs that showed excessive redundancy with ASVAB constructs--again, so that steps could be taken to alleviate such problems for the Pilot Trial Battery. Correlation matrices and factor analyses were the major methods of analysis for these second and third purposes.

The psychometric analyses showed some problems with the cognitive tests. The time limits appeared too stringent for several tests, and one test, Hidden Figures, appeared to be much too difficult for the population being tested. Since most of the cognitive tests used in the Preliminary Battery had been developed on college samples or other samples somewhat better educated than the population seeking entry into the Army, these findings were not unexpected. The lesson learned was that the Pilot Trial Battery measures needed to be accurately targeted (in difficulty of items and time limits) toward the population of persons seeking entry into the Army. No serious problems were unearthed for the temperament, biodata, and interest scales. Item-total correlations were acceptably high and in accordance with prior findings, and score distributions were not excessively skewed or different from expectation. About 8% of subjects failed the scale that screened for inattentive or random responding on the temperament inventory, a figure that is in accord with findings in other selection research.

Covariance analyses showed that vocational interest scales were relatively distinct from the biographical and temperament scales, but the latter two types of scales showed considerable covariance. Five factors were identified from the 40 non-cognitive scales, two that were primarily vocational interests and three that were combinations of biographical data and temperament scales. These findings led us to consider, for the Pilot Trial Battery, combining biographical and temperament item types to measure the constructs in these two areas. The five non-cognitive factors showed relative independence from the cognitive PB tests, with the median absolute correlations of the scales within each of the five factors with each of the eight PB cognitive tests ranging from .01 to .21. This confirmed our expectations of little or no overlap between the cognitive and non-cognitive constructs.

Correlations and factor analysis of the ten ASVAB subtests and the eight PB cognitive tests confirmed prior analyses of the ASVAB (Kass, et al., 1983) and the relative independence of the PB tests. Although some of the ASVAB-PB test correlations were fairly high (the highest was .57), most were less than .30 (49 of the 80 correlations were .30 or less, 65 were .40 or less). The factor analysis (principal factors extraction, varimax rotation) of the 18 tests showed all eight PB cognitive tests loading highest on a single factor, with none of the ASVAB subtests loading highest on that factor. The non-cognitive scales overlapped very little with the four ASVAB factors identified in the factor analysis of the ASVAB subtests and PB cognitive tests. Median correlations of non-cognitive scales with the ASVAB factors, computed within the five non-cognitive factors, ranged from .03 to .32, but 14 of the 20 median correlations were .10 or less.

COMPUTER BATTERY DEVELOPMENT

Roughly speaking, four phases of activities led up to the development of computerized predictor measures for the Pilot Trial Battery: (1) information gathering about past and current research in perceptual/psychomotor measurement and computerized methods of testing such abilities; (2) construction of a demonstration computer battery, and a continuation of information gathering; (3) selection of commercially available microprocessors and peripheral devices, writing of software for testing several abilities using this hardware, and try out of this hardware and software; (4) continued development of software, and design and construction of a custom-made peripheral device, which we called a response pedestal.

Background

Compared to the paper-and-pencil measurement of cognitive abilities and the major non-cognitive variables (temperament, biographical data, and vocational interests), the computerized measurement of psychomotor and perceptual abilities was in a relatively primitive state of knowledge. Much work had been done in World War II using electro-mechanical apparatus, but relatively little work had occurred since then. Microprocessor technology held out the promise of revolutionizing measurement in this area, but the work was (and still is) in its early stages. It was clear, however, that cognitive ability testing was moving into a computer-assisted environment through the methodology of adaptive testing. As Project A began, work was under way to implement the ASVAB via computer-assisted testing methods in the Military Entrance Processing Stations. Therefore, it was also sensible from a practical point of view to investigate these methods of testing.

It was with this backdrop of relatively little research-based knowledge, excitement at the prospect of microprocessor-driven and, therefore, accurate and reliable testing, and the looming implementation of computerized testing in the military environment, that we began our work.

Phase 1. Information Gathering

The two major activities in this phase were literature review and visits to several military laboratories that were engaged in apparatus, simulator, or microprocessor-driven testing of psychomotor and other abilities.

The literature review procedures were described earlier. Almost no literature was available on computerized, especially microprocessor-driven, testing of psychomotor/perceptual abilities for selection/classification purposes. Considerable literature was available on the taxonomy or structure of such abilities, based primarily on work done in World War II or shortly thereafter. Work from this era showed that testing such abilities with electro-mechanical apparatus did show useful levels of validity for such jobs as aircraft pilot, but that such apparatus had reliability problems. This information focused our attention on the types of abilities that would provide an efficient, yet comprehensive, coverage of this ability domain, confirmed the notion that testing such abilities could yield useful validities, but emphasized the problems with unreliability in the use of electro-mechanical apparatus.

To obtain the most current information, in the spring of 1983 we visited four military laboratories engaged in relevant research: the Air Force Human Resources Laboratory (AFHRL), Brooks Air Force Base; the Naval Aerospace Medical Research Laboratory (NAMRL), Pensacola Naval Air Station; the Army Research Institute Field Unit at Fort Rucker, Alabama; and the Army Research Institute Field Unit at Fort Knox, Kentucky. We were primarily after the answers to five questions:

1. What computerized measures are in use?

   We found more than sixty different measures in use across the four sites. (Appendix E shows the names, location, and associated hardware/software for these measures.) A sizable number were specialized simulators that were not relevant for Project A (e.g., a helicopter simulator weighing several tons that is permanently mounted in an air-conditioned building). However, many measures in the perceptual, cognitive, and psychomotor areas were relevant.

2. What computers were selected for use? and,

3. What computer languages are being used?

   We observed three different microprocessors in use--the Apple, Terak, and PDP 11--and three different computer languages--PASCAL, BASIC, and FORTRAN. There appeared to be relatively little in common among the four sites in terms of the hardware/software used.

4. How reliable are these computerized measures? and,

5. What criterion-related validity evidence exists for these measures so far?

   Data were currently being collected at all four sites to address the reliability and criterion-related validity questions, but very little documented information was available. The research at AFHRL was at the point of administering computerized measures to fairly large samples of subjects. This was also true of the research at Fort Rucker, where they expected to have validity data collected and analyzed by sometime in 1984.

   A number of the measures had been under study at NAMRL for some time, but criterion-related validity had not been the primary focus of their work. The prototype information processing measures developed there had been shown to be sensitive to individual differences within chronological age groups as well as to age-related changes across different age groups. We were not able to observe these measures directly as they were being administered off-site, under NAMRL contract at the Aviation Research Laboratory in Illinois, but the research was described to us in some detail.

Data on the computerized measures at Fort Knox were being analyzed. Their efforts apparently were hampered by severe range restriction in the predictors as well as some problems with the criterion measures. They were finding significant, positive correlations between microprocessor measures and their higher fidelity, "hands-on" counterparts.

To summarize, little information was then available on the reliability or criterion-related validity of the computerized measures in use at the sites. This was not surprising since most of the measures had been developed only recently.

Nevertheless, we learned some valuable lessons. First, large-scale testing can be carried out on microprocessor equipment (AFHRL was doing so). Second, a variety of software and hardware can produce satisfactory results, but we should carefully evaluate options before making these choices. Third, it would be highly desirable to have the testing devices or apparatus be as compact and simple in design as possible, in order to minimize "down" time and make transportation feasible. Fourth, we began to form the impression that it would be highly desirable to develop our software and hardware devices to be as completely self-administering (i.e., little or no input required from test monitors) as possible and as impervious as possible to prior experience with typewriting and playing video games.

Phase 2. Demonstration Battery

After conducting the site visits, we programmed a short demonstration battery in the BASIC language on the Osborne 1, a portable microprocessor. The purpose was to implement some of the techniques and procedures observed during the visits in order to determine the degree of difficulty of such programming, and to get an idea of the quality of results to be expected from using a common portable microprocessor and a language that is common to many machines but has some disadvantages in processing power, speed, and flexibility.

This short battery was self-administering, recorded time-to-answer and the answer made, and contained five tests: simple reaction time (pressing a key when a stimulus appeared), choice reaction time (pressing one of two keys in response to one of two stimuli), perceptual speed and accuracy (comparing two alphanumeric phrases for similarity), verbal comprehension (vocabulary knowledge), and a self-rating form (indicating which of two adjectives "best " describes the examinee, on a 7-point scale). We also experimented with the programming of several types of visual tracking tests, but did not include these in the self-administered demonstration battery.

No data were collected with this demonstration battery, but it fulfilled its intended purposes. Experience in developing and using the battery convinced us that the BASIC language did not allow enough power and control of timing events to be useful for our purposes. The basic methods for controlling stimulus presentation and response acquisition through a keyboard were thoroughly explored. Techniques for developing a self-administering battery of tests were tried out.

The second activity during this phase was consultation at the University of Illinois with three experts about perceptual/psychomotor abilities and their measurement.[1] We met with them to review what we had learned from our activities to date, discuss our near-term development plans, and get their reactions. We also discussed their program of research in this area and observed their computerized testing facility. The major points that emerged from this meeting were:

- Generally speaking, it may be difficult to obtain discriminant validity with the addition of new predictors (beyond the ASVAB), but the approach being taken by Project A Task 2 seems to allow the maximal opportunity for this to occur and it allows the testing of the hypothesis.

- The results obtained in World War II using electro-mechanical, psychomotor testing apparatus probably do generalize to the present era in terms of the structure of abilities and the usefulness of such abilities for predicting job performance in jobs like aircraft pilot.

- The taxonomy of psychomotor skills and abilities probably should be viewed in a hierarchical fashion, and perhaps Project A's development efforts would be best focused on two or three relatively high-level abilities such as gross motor coordination, multilimb constant processing tasks, and fine manipulative dexterity.

- Rate of learning or practice effects are viewed as a major concern for evaluating the usefulness of psychomotor ability measures for predicting on-the-job performance. If later test performance (after many trials) was much more valid than early test performance (early trials), or worse, if early test performance was not valid and later test performance was, then it is unlikely that psychomotor testing would be practically feasible in the operational military-selection environment. There are, however, no empirically based answers to these questions, and it is acknowledged that research is necessary to obtain answers, especially with microprocessor-driven testing methods.

Phase 3. Selection/Purchase of Microprocessors and Development/Tryout of Software

On the basis of the information from the first two phases, we defined the desirable characteristics of a microprocessor useful for our research. A prime consideration was transportability. Almost all of our pilot testing and other data collection efforts would take place at various field sites throughout the United States and Europe. We would not be able to build a stationary laboratory and bring the soldiers to the site.

Following are the desired characteristics as we outlined them in the Fall of 1983:

---

[1] Charles Hulin, John Adams, and Phillip Ackerman were the consultants.

1. Reliability--This encompasses several considerations. First, the machine should be manufactured and maintained by a company that has a history of backing its products and, even more basic, is likely to remain in business. Second, the machine itself should be fairly rugged and capable of being carried around without breaking down.

2. Portability--Since we will need to transport the computer to several posts during development efforts, the machine should be as portable as possible, and, if feasible, extremely easy to assemble and disassemble.

3. Most Recent Generation of Machine--Progress is very rapid in this area; therefore, we should get the latest "proven" type of machine. That means getting a 16-bit microprocessor rather than an 8-bit microprocessor. This way, the software developed will be more likely to be usable on future machines.

4. Compatibility--Although extremely difficult to achieve, a desirable goal is to have a machine that is maximally compatible with other machines, or that will have software that will be compatible with other machines. Thus, we think a CPM-based machine or some version of the 8088 chip is best.

5. Appropriate Display Size, Memory Size, Disk Drives, Graphics, and Peripheral Capabilities--We need a video display that is at least nine inches (diagonally), but it need not be a color monitor. Since we will be developing experimental software, we need a relatively large amount of random access memory, and 256 K seems to be the largest memory size that is generally available. (Later project efforts to create maximally efficient use of memory may considerably reduce this requirement.) Also we require two floppy disk drives to store needed software and to record subjects' responses. High-resolution graphics capability is desirable for some of the kinds of tests we will develop. Finally, since several of the ability measurement processes will require the use of paddles, joysticks, or other similar devices, the machine must have the appropriate hardware and software to allow this.

The characteristics listed in the above statement were used as criteria for evaluating commercially available microprocessors. Most machines were eliminated because they were very new on the market and thus had no history, or they were made by relatively unknown manufacturers.

In the end we selected Compaq portable microprocessors with 256 K random access memory, two 320 K-byte disk drives, a "game board" for accepting input from peripheral devices such as joysticks, and software for FORTRAN, PASCAL, BASIC, and assembly language programming. Six of these machines were purchased in December 1983. We also purchased six commercially available, dual-axis joysticks.

We then developed the initial version of the software needed to test several perceptual/psychomotor abilities that we were reasonably certain would be chosen for final inclusion in the Pilot Trial Battery, although those abilities had not yet been finally selected. We had three general, operational objectives in mind for the software to be produced: (1) as far as possible, it should be transportable to other microprocessors; (2) it should require as little intervention as possible from a test administrator in the process of presenting the tests to subjects and storing the data; and (3) it should enhance the standardization of testing by adjusting for hardware differences across computers and response pedestals.

We first had to choose a primary language. We chose to prepare the bulk of the software using the PASCAL language as implemented by Microsoft, Inc. PASCAL is a common language and it is implemented using a compiler that permits modularized development and software libraries. As computer languages go, PASCAL is relatively easy for others to read and it can be implemented on a variety of computers.

Some processes, mostly those that are specific to the hardware configuration, had to be written in IBM-PC assembly language. Examples include interpretation of the peripheral device inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a PASCAL-callable "primitive" routine with a unitary purpose was written in assembly language. Although the machine-specific code would be useless on a different type of machine, the functions were sufficiently simple and unitary in purpose so that they could be reproduced with relative ease.

The overall strategy of the software development was to take advantage of each researcher's input as directly as possible. It quickly became clear that the direct programming of every item in every test by one person (a programmer) was not going to be very successful in terms of either time constraints or quality of product. To make it possible for each researcher to contribute his/her judgment and effort to the project, it was necessary to plan so as to, as much as possible, take the "programmer" out of the step between conception and product and enable researchers to create and enter items without having to know special programming.

The testing software modules were designed as "command processors" which interpreted relatively simple and problem-oriented commands. These were organized in ordinary text written by the various researchers using word processors. Many of the commands were common across all tests. For instance, there were commands that permitted writing of specified text to "windows" on the screen and controlling the screen attributes (brightness, background shade, etc); a command could hold a display on the screen for a period measured to 1/100th-second accuracy. There were commands that caused the program to wait for the respondent to push a particular button. Other commands caused the cursor to disappear or the screen to go blank during the construction of a complex display.

Some of the commands were specific to particular item types. These commands were selected and programmed according to the needs of a particular test type. For each item type, we decided upon the relevant stimulus properties to vary and built a command that would allow the item writer to quickly construct a set of commands for items which he or she could then inspect on the screen.

These techniques made it possible for entire tests to be constructed and experimentally manipulated by psychologists who could not program a computer.

As this software was written, we used it to administer the computerized tests to small groups of soldiers (N = 5 or fewer) at the Minneapolis Military Entrance Processing Station (MEPS). These soldiers were told about Project A, that their participation was voluntary and the test results would not affect their status, but that we needed to have them try their very best so that we could evaluate the tests. They were also asked to write down anything about the tests that bothered them or any problems they encountered during the testing, and told that the researchers would talk to them about the computerized test battery when they were finished. The soldiers completed the battery without assistance from the researchers, unless it was absolutely necessary, and were then questioned.

The nature of these questions varied over the progress of these developmental tryouts, but mainly dealt with clarity of instructions, difficulty of tests or test items, screen brightness problems, difficulties using keyboard or joysticks, clarity of visual displays, and their general (favorable/unfavorable) reaction to this type of testing.

These tryouts were held from 20 January 1984 through 1 March 1984, and a total of 42 persons participated in nine sessions. The feedback from the participants was extremely useful in determining the shape of the tests, prior to the first pilot test of the Pilot Trial Battery. After each tryout, we would modify the software to clarify instructions, make item or test difficulties more appropriate, make stimulus displays and sequences of events more appropriate, and so forth. We also performed simple analyses of the data collected, but mainly to insure that responses were being captured and recorded correctly--not for any substantive analyses of the tests or constructs.

At the end of Phase 3, we had developed a self-administering, computerized test battery that was implemented on a Compaq portable computer. The subjects responded on the normal keyboard for all tests except a tracking test that required them to use a joystick. This joystick was a commercially available device normally used for video games. Seven different tests had been programmed. These were not necessarily tests we wished to include in the Pilot Trial Battery, but five did eventually end up in that battery.

Phase 4. Continued Software Development and Design/Construction of a Response Pedestal

During the fourth phase, several significant events took place during March-May 1984. An in-progress-review (IPR) meeting was held at which we presented the results of the development efforts to date and received guidance on next efforts from ARI staff, the Scientific Advisory Group subcommittee assigned to Task 2, and other Project A researchers. We made field observations of some combat MOS in order to inform the further development of computerized tests; the first pilot test of the computerized battery was completed; and we designed and constructed a custom-made response pedestal for the computerized battery.

The primary result of the in-progress-review was the identification and prioritization of the ability constructs for which computerized tests should be developed. Chapter 5 describes these constructs in some detail.

A second result of the review was a decision to go to the field to observe several combat arms MOS in order to target the tests more closely to those skills, insofar as that was possible.

These field observations subsequently took place at several posts. They were relatively informal; we simply observed soldiers (usually a very small number) working at their jobs in the field and, where possible, asked questions to clarify their activities. We did complete a brief checklist that required a rating of the degree of importance for the job of several cognitive, perceptual, and psychomotor abilities; these checklists were not formally analyzed but were used for later discussions and development efforts. We also operated various training aids and simulators available during our visits. The MOS for which we were able to complete these field observations were: 11B (Infantryman), 13B (Cannon Crewman), 19K (Armor Crewman), 16S (MANPADS Crewman), and 05C (Radio Teletype Operator).

On one of these site visits we were able to administer the computerized battery to several trainers (for Armor Crewman, 19K). The primary outcome of their feedback was a decision to develop a test that utilized military aircraft and vehicle profiles in an identification task. Their suggestion corroborated our field observations that such a test seemed more appropriate than a test then in the battery that was intended to predict skill at target identification (this test had been adapted from the Hidden Figures test in the ETS battery).

The first pilot test of the Pilot Trial Battery occurred at Fort Carson during this phase. (See Chapter 2 for a description of the sample and procedures of that pilot test.) For the computerized tests, the same procedures were used as for the MEPS tryouts described above in Phase 3. A total of 20 soldiers completed the computerized battery.

The information from this pilot test primarily confirmed a major concern that had surfaced during the MEPS tryouts, namely the undesirability of the computer keyboard and commercially available joysticks for acquiring test responses. Feedback from subjects (and our observations) indicated that (1) it was difficult to pick out one or two keys on the keyboard, and (2) fairly elaborate, and therefore confusing, instructions were needed to use the keyboard in this manner. Even with such instructions, subjects often missed the appropriate key, or inadvertently pressed the keys because they were leaving their fingers on the key in order to retain the appropriate position for response. Also, subjects varied in the way they prepared for test items, and the more or less random positioning of their hands added unwanted (error) variance to their scores.

Similar issues arose with the joysticks, but the main problems were their lack of durability and the large variance across joysticks in their operating characteristics, again adding error variance.

After consultation with ARI and other Project A researchers, we decided to develop a custom-made response pedestal in an attempt to alleviate these problems. We drew up a rough design for such a pedestal and contracted with an engineering firm to fabricate a prototype. We tried out the first prototype, suggested modifications, and had six copies produced in time for the Fort Lewis pilot test in June 1984. Chapter 5 describes the response pedestal in some detail.

Completing work in Phase 4 we wrote additional software to (1) test the abilities that had been chosen for inclusion in the Pilot Trial Battery and (2) accommodate the new response pedestal.

PILOT TRIAL BATTERY

Identification of Measures

In March 1984, an IPR meeting was held to decide on the measures to be developed for the Pilot Trial Battery. Information from the literature review, expert judgments, initial analyses of the preliminary battery, and the first three phases of computer battery development was presented and discussed. Task 2 staff made recommendations for inclusions of measures and these were evaluated and revised. Figure 1.5 shows the results of that deliberation process. (The names of the tests developed for the Pilot Trial Battery are shown in the right-hand column of Figure 1.5. Each of these tests is dealt with extensively in later chapters, so we make no attempt to describe them here.) This set of recommendations served as the blueprint for Task 2's test development efforts for the next several months.

Pilot Tests and Field Tests

There were three pilot tests of the measures developed for the Pilot Trial Battery. These took place at Fort Carson in April 1984, Fort Campbell in May 1984, and Fort Lewis in June 1984. At the first two sites not all Pilot Trial battery measures were administered, but the complete battery was administered at Fort Lewis. Subsequent chapters of this report describe these pilot tests, resulting analyses, and revisions to measures prior to the field tests. The reports of analyses of the pilot test data emphasize the Fort Lewis administration because it was the first time the complete battery was administered and it was the largest pilot test sample. (The pilot tests, especially those at Fort Carson and Fort Campbell, are often referred to as "tryouts" in the remainder of this report.)

A field test of the complete Pilot Trial Battery was conducted at Fort Knox in September 1984. In addition, supplementary field test studies were conducted at Fort Knox, Fort Bragg, and the Minneapolis MEPS during the Fall of 1984. Following analysis of the field test results, the test battery was revised for use in the Concurrent Validation administration.

The data collection procedures and samples for the various tests are described in Chapter 2 of this report. Description of the measures themselves, and of the results of the tests and analyses, is organized by the major types of predictor categories:

Cognitive Paper-and-Pencil	Chapter 3, Pilot Tests, and Chapter 4, Field Test
Perceptual/Psychomotor, Computer-Administered	Chapter 5, Pilot Tests, and Chapter 6, Field Test
Non-Cognitive Paper-and-Pencil	Chapter 7, Pilot Tests, and Chapter 8, Field Test

Revisions of the measures after field testing, into the form to be used in Concurrent Validation, are described in Chapter 9.

Final Priority*	Predictor Category	Pilot Trial Battery Test Names
Cognitive:		
7	Memory	(Short) Memory Test - Computer
6	Number	Number Memory Test - Computer
8	Perceptual Speed & Accuracy	Perceptual Speed & Accuracy - Computer Target Identification Test - Computer
4	Induction	Reasoning Test 1 Reasoning Test 2
5	Reaction Time	Simple Reaction Time - Computer Choice Reaction Time - Computer
3	Spatial Orientation	Orientation Test 1 Orientation Test 2 Orientation Test 3
2	Spatial Visualization/Field Independence	Shapes Test
1	Spatial Visualization	Object Rotations Test Assembling Objects Test Path Test Maze Test
Non-Cognitive, Biodata/Temperament:		
1	Adjustment	ABLE (Assessment of Background Life Experiences)
2	Dependability	
3	Achievement	
4	Physical Condition	
5	Potency	
6	Locus of Control	
7	Agreeableness/Likeability	
1	Validity Scales	
Non-Cognitive, Interests:		
1	Realistic	AVOICE (Army Vocational Interest Career Examination)
2	Investigative	
3	Conventional	
4	Social	
5	Artistic	
6	Enterprising	
Psychomotor:		
1	Multilimb Combination	Target Tracking Test 2 - Computer Target Shoot - Computer
2	Precision	Target Tracking Test 1 - Computer
3	Manual Dexterity	(None)

*Final priority arrived at via consensus of March 1984 IPR attendants.

Figure 1.5. Predictor categories discussed at IPR March 1984, linked to Pilot Trial Battery test names

Chapter 1 References

Alley, W. E., & Matthews, M. D. (1982). The vocational interest career examination. *Journal of Psychology, 112*, 169-193.

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). Developing a model of soldier effectiveness: A strategy and preliminary results. Presented at the 91st Annual Convention of the American Psychological Association, Anaheim, California. In Eaton, N.K., & Goer, M. H. (Eds.) (1983). *Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report* (ARI Research Note 83-37).

Bownas, D. A., & Heckman, R. W. (1976). *Job analysis of the entry-level firefighter position.* Minneapolis, MN: Personnel Decisions, Inc.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Flanagan, J. C. (1965). *Flanagan industrial test manual.* Chicago: Science Research Associates.

Gough, H. G. (1975). *Manual for the California Psychological Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J. S., & Peterson, N. G. (1984). Covariance analyses of cognitive and non-cognitive measures of Army recruits: An initial sample of Preliminary Battery Data. Presented at the 92nd Annual Convention of the American Psychological Association, Toronto. In Eaton et al. (Eds.) (1984). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year* (ARI Technical Report 660). Alexandria, VA: Army Research Institute.

Hough, L. M., Kamp, J. D., & Barge, B. N. *Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance.* ARI Research Note in preparation.

Jackson, D. N. (1967). *Personality Research Form Manual.* Goshen, NY: Research Psychologists Press.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10: 1981 Army applicant sample. *Educational and Psychological Measurement, 43*, 1077-1088.

McHenry, J. J., & Rose, S. R. *Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification.* ARI Research Note in preparation.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology Monographs, 64*, 569-607.

Peterson, N. G., & Bownas, D. A. (1982). Skills, task structure, and performance acquisition. In Marvin D. Dunnette and Edwin A. Fleishman (Eds.) *Human performance and productivity* (Vol. I). Hillsdale, NJ: Lawrence Erlbaum Associates.

Peterson, N. G., & Houston, J. S. (1980). *The prediction of correctional officer job performance: Construct validation in an employment setting.* Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., Bosshardt, M. J., & Dunnette, M. D. (1977). *A study of the correctional officer job at Marion Correctional Institution, Ohio: Development of selection procedures, training recommendations and an exit information program.* Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., & Rosse, R. L. (1984). *The LOMA job effectiveness prediction system: Validity analyses* (Technical Report No. 4). Atlanta, GA: Life Office Management Association.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80*, (1, Whole No. 609).

Ruch, F. L., & Ruch, W. W. (1980). *Employee Aptitude Survey: Technical Report.* Los Angeles, CA: Psychological Services, Inc.

Schmidt, F. L., Hunter, J. E., Croll, P. R., & KcKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. *Journal of Applied Psychology, 68*, 590-601.

Tellegen, A. (1982). *Brief manual for the differential personality questionnaire.* Unpublished manuscript, University of Minnesota.

Toquam, J. L., Corpe, V. A., & Dunnette, M. D. *Literature review: Cognitive abilities--theory, history, and validity.* ARI Research Note in preparation.

Wing, H., Peterson, N. G., & Hoffman, R. E. (1984). *Expert judgments of predictor-criterion validity relationships.* Presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. In Eaton et al. (Eds.) (1984). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year* (ARI Tehcnical Report 660). Alexandria, VA: Army Research Institute.

CHAPTER 2

TEST DATA COLLECTION: PROCEDURES AND SAMPLES

Janis S. Houston

In this chapter, we describe the procedures used to collect data at the pilot and field test sites and report basic descriptive data about the sample of soldiers that participated.

PILOT TESTS

Pilot Test #1: Fort Carson

Procedures

On 17 April 1984, a sample of 43 soldiers at Fort Carson, Colorado participated in the first pilot testing of the Pilot Trial Battery. The testing session ran from 0800 hours to 1700 hours, with two 15-minute breaks (one mid-morning and one mid-afternoon) and a one-hour break for lunch.

Groups of five soldiers at a time were randomly selected to take computerized measures in a separate room while the remaining soldiers took paper-and-pencil tests. When a group of five soldiers completed the computerized measures, they were individually and collectively interviewed about their reactions to the computerized tests, especially regarding clarity of instructions, face validity of tests, sensitivity of items, and their general disposition toward such tests. The soldiers then returned to the paper-and-pencil testing session, and another group of five was selected to take the computer measures.

Thus, not all the soldiers took all of the tests. The maximum N for any single paper-and-pencil test was 38 (43 minus the 5 taking computer tests). Computerized measures were administered to a total of 20 soldiers. The new paper-and-pencil cognitive tests in the Pilot Trial Battery were each administered in two equally timed halves, to investigate the Part 1/Part 2 correlations as estimates of test reliability.

After actual test administration was completed, ten soldiers were selected to give specific, test-by-test feedback about paper-and-pencil tests in a small group session, while the remaining soldiers participated in a more general feedback and debriefing session.

Tests Administered

Table 2.1 contains a list of all the tests administered at Fort Carson, in the order in which they were administered, with the time limit and number of items for each test. These tests can be categorized as follows:

- o 10 new cognitive paper-and-pencil measures
- o 9 marker tests for new paper-and-pencil cognitive measures
- o 7 computerized measures

Table 2.1

Pilot Tests Administered at Fort Carson, 17 April 1984

Test	Time Limit (Mins.)	No. of Items	Type of Test
Paper-and-Pencil Tests			
1. Path Test	9	35	New, Cognitive
2. Reasoning Test 1	14	30	New, Cognitive
3. EAS Test 1 - Verbal Comprehension	5	30	Marker, Cognitive
4. Orientation Test 1	8	20	New, Cognitive
5. Shapes Test	16	54	New, Cognitive
6. EAS Test 2 - Numerical Ability	10	75	Marker, Cognitive
7. Object Rotation Test	7	60	New, Cognitive
8. ETS Choosing a Path	8	16	Marker, Cognitive
9. Orientation Test 2	8	20	New, Cognitive
10. Reasoning Test 2	11	32	New, Cognitive
11. Orientation Test 3	12	20	New, Cognitive
12. Assembling Objects Test	16	30	New, Cognitive
13. Maze Test	9	24	New, Cognitive
14. Mental Rotations Test	10	20	Marker, Cognitive
15. ETS Hidden Figures	14	16	Marker, Cognitive
16. ETS Map Planning	6	40	Marker, Cognitive
17. ETS Figure Classification	8	14	Marker, Cognitive
18. EAS Test 5 - Space Visualization	5	50	Marker, Cognitive
19. FIT Assembly	10	20	Marker, Cognitive
Computer Measures[a]			
1. Simple Reaction Time	None	15	New, Perceptual/ Psychomotor
2. Choice Reaction Time	None	15	New, Perceptual/ Psychomotor
3. Perceptual Speed & Accuracy	None	80	New, Perceptual/ Psychomotor
4. Tracing Test	None	26	New, Perceptual/ Psychomotor
5. Short Memory Test	None	50	New, Perceptual/ Psychomotor
6. Hidden Figures Test	None	32	New, Perceptual/ Psychomotor
7. Target Shoot	None	20	New, Perceptual/ Psychomotor

---

[a] All computer measures were administered using a Compaq portable microprocessor with a standard keyboard plus a commercially available dual-axis joystick.

The new paper-and-pencil cognitive tests were tests newly developed by the researchers to measure the constructs or abilities that had been selected as important in earlier stages of the research (see Chapter 1). Detailed descriptions of the development and analyses of these tests are given in Chapters 3 and 4. The marker tests were published tests that were viewed as the closest or best measure of the selected abilities.

Sample Description

As previously mentioned, a total of 43 soldiers participated in Pilot Test #1, with 20 soldiers completing the computerized measures and a maximum of 38 soldiers completing individual paper-and-pencil tests. Table 2.2 presents a brief demographic description of the sample.

Table 2.2

Description of Fort Carson Sample (N = 43)

1. Age:

Mean = 22.76 years

Median = 21.50 years

Standard Deviation = 2.19

2. Current MOS:

MOS	N
19	8
11	6
13	5
16	4
98	3
05	2
27	2
64	2
76	2
91	2
96	2
24	1
31	1
36	1
71	1
75	1

3. Sex:

Males	33
Females	10

4. Race:

Black	10
Asian	1
White	24
Hispanic	5
Other	3

5. Years in the Service:
(Computed from Date of Enlistment)

Mean = 1.72

Median = 1.55

Standard Deviation = 1.10

Pilot Test #2: Fort Campbell

Procedures

The second pilot testing session was conducted at Ft. Campbell, Kentucky on 16 May 1984. A sample of 57 soldiers attended the 8-hour session, and all 57 completed paper-and-pencil tests. No computerized measures were administered at this pilot session. Once again, the ten new cognitive tests were administered in two equally timed halves, to investigate Part 1/Part 2 correlations.

Because we were still experimenting with time limits on the new cognitive tests, soldiers were asked to mark which item they were on when time was called for each of these tests, and then to continue to work on that part of the test until they finished. Finishing times were recorded for all the tests (Parts 1 and 2 separately, where appropriate).

After test administration was completed, the group was divided. Ten individuals were selected to provide specific feedback concerning the new non-cognitive measures and the remaining individuals provided feedback on the new cognitive measures.

Tests Administered

Table 2.3 lists all the tests and inventories administered at Pilot Test #2: Fort Campbell, along with the time limit and number of items for each. There were ten new cognitive tests with five cognitive marker tests, and two new non-cognitive inventories with one non-cognitive marker inventory. No computerized measures were administered.

The two new non-cognitive inventories were developed by the researchers to measure the constructs selected as important in earlier stages of the research (see Chapter 1). The Assessment of Background and Life Experiences (ABLE) measured temperament and biodata constructs and the Army Vocational Interest Career Examination (AVOICE) measured vocational interests. The Personal Opinion Inventory (POI) was intended as a marker inventory in that it contained published scales thought to measure the constructs selected as important in the temperament domain. Detailed descriptions of the rationale, development, and analyses of the new non-cognitive inventories are provided in Chapters 7 and 8.

Sample Description

A total of 57 soldiers completed the Pilot Trial Battery as administered at Fort Campbell. A description of this sample's demographic make-up appears in Table 2.4.

Pilot Test #3: Fort Lewis

Procedures

For the third pilot testing session, approximately 24 soldiers per day for five days (11-15 June 1984) were available for testing at Fort Lewis, Washington. Test sessions ran from 0800 hours to 1700 hours, with short breaks in the morning and afternoon, and a one-hour lunch break. The

Table 2.3

Pilot Tests Administered at Fort Campbell, 16 May 1984

Paper-and-Pencil Tests	Total Time Limit (Mins.)	No. of Items	Type of Test
1. Path Test	9	44	New, Cognitive
2. Reasoning Test 1	14	30	New, Cognitive
3. EAS Verbal Comprehension	5	30	Marker, Cognitive
4. Orientation Test 1	9	30	New, Cognitive
5. Shapes Test	16	54	New, Cognitive
6. Object Rotation Test	9	90	New, Cognitive
7. Reasoning Test 2	11	32	New, Cognitive
8. Orientation Test 2	8	20	New, Cognitive
9. ABLE (Assessment of Background and Life Experiences)	None	291	New, Non-Cognitive
10. Orientation Test 3	12	20	New, Cognitive
11. Assembling Objects Test	16	40	New, Cognitive
12. Maze Test	8	24	New, Cognitive
13. AVOICE (Army Vocational Interest Career Examination)	None	306	New, Non-Cognitive
14. ETS Hidden Figures	14	16	Marker, Cognitive
15. ETS Map Planning	6	40	Marker, Cognitive
16. ETS Figure Classification	8	14	Marker, Cognitive
17. FIT Assembly	10	20	Marker, Cognitive
18. POI (Personal Opinion Inventory)	None	121	Marker, Non-Cognitive

Table 2.4

Description of Fort Campbell Sample (N = 57)

1. Age:

Mean = 21.40 years
Median = 21 years
SD = 3.07

2. Current MOS:

MOS	N
76	19
63	11
27	9
52	9
31	3
36	2
71	2
54	1
62	1

3. Sex:

Males	46
Females	11

4. Race:

Black	15
Asian	1
White	36
Hispanic	5

5. Years in the Service:
(Computed from Date of Enlistment)

Mean = 1.84

Median = 1.67

SD = 1.27

entire Pilot Trial Battery, including new cognitive and non-cognitive measures, was administered to all soldiers. To accomplish this, the schedule displayed in Table 2.5 was followed.

Each day, the approximately 24 soldiers were divided into four groups (labeled A, B, C, and D) of six soldiers each. While Group A took the computerized measures, groups B, C, and D took the first half of the paper-and-pencil cognitive tests (labeled C1). While Group B took the computerized measures, Groups A, C, and D took the second half of the paper-and-pencil cognitive measures (labeled C2), and while Group C took the computerized measures, Groups A, B, and D took the paper-and-pencil non-cognitive measures (labeled NC). At approximately 1500 hours, each group took that portion of the Pilot Trial Battery they had not yet received.

Once again, the new paper-and-pencil cognitive tests were administered in two equally timed halves to investigate Part 1/Part 2 correlations as estimates of test reliability. Individuals were not allowed any extra time to work on each test beyond the time limits, but finishing times were recorded for individuals completing tests before time was called.

After a soldier completed the computerized battery, each was asked about his or her general reaction to the computerized battery, the clarity and completeness of the instructions, perceived difficulty of the tests, and ease of use of the response apparatus.

Table 2.5

Daily Schedule for Fort Lewis Pilot Testing[a]

Approximate Time	Room 1	Room 2	Room 3
0800 to 0815	A, B, C, D for Introduction, etc.		--
0815 to 1000	B, C, D take first half of Cognitive Tests (C1)	A takes all computer measures	--
1015 to 1200	A, C, D take second half of Cognitive Tests (C2)	B takes all computer measures	--
1300 to 1500	A, B, D take all Non-Cognitive Measures (NC)	C takes all computer measures	--
1515 to 1700	A takes C1	D takes all computer measures C takes NC	B takes C2

[a] Each day the soldiers in the sample were divided into four groups of approximately six soldiers each, referred to here as Groups A, B, C, and D.

Tests Administered

The tests administered at Pilot Test #3 in Fort Lewis, are listed in Table 2.6, with the time limit and number of items in each test. A summary of these tests follows:

- 10 new, paper-and-pencil, cognitive tests
- 4 marker, paper-and-pencil, cognitive tests
- 2 new, paper-and-pencil, non-cognitive tests
- 8 new, computerized, perceptual/psychomotor measures

Sample Description

Table 2.7 provides demographic information about the Fort Lewis sample. A total of 118 soldiers participated in the pilot testing.

Summary of Pilot Tests

The Pilot Test Battery was initially developed in March 1984 and went through three complete pilot testing iterations by August 1984. After each iteration, observations noted during administration were scrutinized, data analyzed, and results carefully examined. Revisions were made in specific item content, test length, and time limits, where appropriate.

Table 2.8 summarizes the three Pilot Test sessions conducted during this period, with the total sample size for each, and the number and types

Table 2.6

Pilot Tests Administered at Fort Lewis, 11-15 June 1984

Administration Group	Test	Total Time Limit	No. of Items	Type of Test
	Paper-and Pencil Tests			
C1	Path Test	8	44	New, Cognitive
	Reasoning Test 1	12	30	New, Cognitive
	Orientation Test 1	10	30	New, Cognitive
	Shapes Test	16	54	New, Cognitive
	Object Rotation Test	8	90	New, Cognitive
	Reasoning Test 2	10	32	New, Cognitive
	Maze Test	6	24	New, Cognitive
C2	SRA Word Grouping	5	30	Marker, Cognitive
	Orientation Test 2	10	24	New, Cognitive
	Orientation Test 3	12	20	New, Cognitive
	Assembling Objects Test	16	40	New, Cognitive
	ETS Map Planning	6	40	Marker, Cognitive
	Mental Rotations Test	10	20	Marker, Cognitive
	DAT Abstract Reasoning	13	25	Marker, Cognitive
NC	ABLE	None	268	New, Non-Cognitive
	AVOICE	None	306	New, Non-Cognitive
	Computerized Measures[a]:			
	Simple Reaction Time	None	15	New, Perceptual/ Psychomotor
	Choice Reaction Time	None	15	New, Perceptual/ Psychomotor
	Perceptual Speed & Accuracy	None	80	New, Perceptual/ Psychomotor
	Target Tracking Test 1	None	18	New, Perceptual/ Psychomotor
	Target Tracking Test 2	None	18	New, Perceptual/ Psychomotor
	Target Identification Test	None	44	New, Perceptual/ Psychomotor
	Memory Test	None	50	New, Perceptual/ Psychomotor
	Target (Shoot) Test	None	40	New, Perceptual/ Psychomotor

---

[a] All computer measures were administered via a custom-made response pedestal designed specifically for this purpose. No responses were made on the computer keyboard. A Compaq microprocessor was used.

Table 2.7

Description of Fort Lewis Sample (N = 118)

1. Age:

Mean = 22.82 years

Median = 22.21 years

SD = 4.2

2. Current MOS:

MOS	N
05B	2
05C	5
11B	13
11C	6
11H	12
13C	1
13E	2
13F	2
19E	1
27E	1
31E	1
31V	3
36C	3
36K	1
54C	5
54E	2
63B	4
63J	1
63W	1
64C	5
67V	4
67Y	2
68G	1
68J	1
71L	4
72E	2
73C	1
74D	1
74F	1
75B	3
75F	1
76C	2
76P	1
76V	5
76W	2
76Y	6
82C	2
83F	1
91B	5
94B	2

3. Sex:

Males	97
Females	22

4. Race:

Black	30
Hispanic	14
White	66
Asian	3
North American Indian	2
Other	1
Blank	2

5. Years in the Service:
(Computed from Date of Enlistment)

Mean = 2.55

Median = 1.75

SD = 2.90

Table 2.8

Summary of Pilot Testing Sessions for Pilot Trial Battery

Pilot Test #	Location	Date	Total Sample Size	No./Type of Tests Administered
1	Fort Carson	17 April 1984	43	10 New Cognitive 9 Marker Cognitive 0 New Non-Cognitive 0 Marker Non-Cognitive 7 Computerized Measures
2	Fort Campbell	16 May 1984	57	10 New Cognitive 5 Marker Cognitive 2 New Non-Cognitive 1 Marker Non-Cognitive 0 Computerized Measures
3	Fort Lewis	11-15 June 1984	118	10 New Cognitive 4 Marker Cognitive 2 New Non-Cognitive 0 Marker Non-Cognitive 8 Computerized Measures

of tests administered at each. Appendix F is a copy of the Pilot Trial Battery as it was administered in June 1984, at Fort Lewis and Appendix G is a copy of the revised Pilot trail Battery as it was administered in the field tests during Fall 1984. (Both Appendix F and Appendix G are contained in a separate limited-distribution report, ARI Research Note 87-24, as noted on page xiv.)

FIELD TESTS

The full Pilot Trial Battery was administered at Fort Knox in September 1984 in a formal field test to evaluate all of the component measures and to analyze psychometric characteristics of the data obtained. In addition, test-retest effects and practice effects were analyzed as part of the Fort Knox field testing, and fakability studies were conducted at Fort Bragg and the Minneapolis Military Entrance Processing Station (MEPS).

Field Test of Pilot Trial Battery: Fort Knox

The field test of the Pilot Trial Battery at Fort Knox was conducted to evaluate the psychometric characteristics of all of the measures in the battery, and to analyze the covariance of the measures with each other and with the ASVAB.

Procedures

Data collection was scheduled for four weeks at Fort Knox. During the first two weeks, 24 soldiers were scheduled each day. On some days, however, more than 24 soldiers arrived for testing. Because of the limited availability of computer testing stations (only six), 24 soldiers was the maximum number that could complete the entire battery. The "overflow" soldiers, however, did complete all of the paper-and-pencil measures.

Each group of soldiers assembled at 0800. The testing sessions included two 15-minute breaks, and one hour was allowed for lunch. When the soldiers were assembled, they were divided into four groups if there were 24 or fewer soldiers, and into five groups if there were more than 24 soldiers.

Figure 2.1 shows the daily schedule of testing for the first two weeks when the full Pilot Trial Battery was being field tested. Figure 2.2 shows the daily schedule in a different way, denoting the room assignments for each group of soldiers throughout the day.

Figure 2.3 shows the schedule for weeks three and four, when the test-retest and practice-effects studies were being conducted. Each soldier from the first two weeks reported back for a half day of testing, either in the morning (0800) or the afternoon (1300), exactly two weeks after his or her week 1 or 2 session. The soldier then completed one-third of all the paper-and-pencil tests (a re-test), and completed either the computer "practice" session or the entire computer battery (a re-test).

Sample Description

If 24 soldiers had appeared for each testing day and completed all tests as scheduled, we would have achieved the following sample sizes:

N = 240 for all cognitive and non-cognitive paper-and-pencil tests

N = 240 for computer tests

N = 80 retest of paper-and-pencil tests

0800 Rollcall. Divide 24 soldiers into four groups of six each, called A, B, C, and D. Overflow soldiers (N>24) were assigned to Group E. (This group's schedule is shown in Figure 2.2).

0815 Read Introduction
Read Privacy Act Statement
Complete Soldier Information Sheet

	Test	Time Limit	
0830	Path Test	8	
	Reasoning Test 1	12	**Cognitive 1 Tests (C1)**
	Orientation Test 1	10	Groups B, C, D complete these. Group A completes computer tests.
	Shapes Test	16	
	Object Rotation Test	7.5	
1030	Reasoning Test 2	10	
	Orientation Test 2	10	**Cognitive 2 Tests (C2)**
	Orientation Test 3	12	Groups A, C, D complete these. Group B completes computer tests.
	Assembling Objects Test	16	
	Maze Test	5.5	
1315	ABLE	50	**Non-Cognitive Instruments (NC)**
	AVOICE	35	Groups A, B, D complete these. Group C completes computer tests.

1515 Final Sessions: Group A takes C1
Group B takes C2
Group C takes NC
Group D takes computer tests

Figure 2.1. Daily testing schedule for Fort Knox Field Test, Weeks 1 and 2.

Approx Time	Room 1	Room 2	Room 3
0800	Assign soldiers to groups: 6 to A, 6 to B, 6 to C, 6 to D, overflow to E. N = 24+		
0815	ABCD for Introduction, Privacy Act & Soldier Info. Sheet N = 24		E for Introduction, Privacy Act, & Soldier Info. Sheet N = overflow, up to 24
0830 to 1015	B, C, D take C1 N = 18	A takes computer tests N = 6	E takes C1
1030 to 1215	A, C, D take C2 N = 18	B takes computer tests N = 6	E takes C2
1315 to 1500	A, B, D take NC N = 18	C takes computer tests N = 6	E takes NC
1515 to 1700	A takes C1 N = 6	D takes computer N = 6 and C takes NC N = 6	B takes C2 N = 6

Figure 2.2. Daily location schedule for Fort Knox Field Test, Weeks 1 and 2.

Daily Schedule for Weeks 3 and 4		
Approx Time	Room 1	Room 2
0800	Week 1: Morning Group A take paper-and-pencil retest* N = 6	Week 1: Morning Group B take computer retest N = 6
1000	Week 1: Morning Group B take paper-and-pencil retest* N = 6	Week 1: Morning Group A take computer practice effects N = 6
1300	Week 1: Aftenoon Group A take paper-and-pencil retest* N = 6	Week 1: Afternoon Group B take computer retest N = 6
1500	Week 1: Afternoon Group B take paper-and-pencil retest* N = 6	Week 1: Afternoon Group A take computer practice effects N = 6

*Each paper-and-pencil retest session received one of the following: C1, C2, or NC. Groups were cycled through all three in that order and the cycle was repeated; i.e., Monday at 0800 is C1, at 1000 is C2, at 1300 is NC, at 1500 is C1; Tuesday at 0800 is C2, etc.

Figure 2.3. Daily schedule for Fort Knox Field Test, Weeks 3 and 4.

N = 120 retest of computer tests

N = 120 practice effects on computer tests

However, due to the usual exigencies of data collection in the field, there was some deviation from these targets. On some days fewer than 24 soldiers appeared, and on other days more than 24 soldiers appeared. In addition, we were able to schedule one additional testing day. Finally, some soldiers were unable to complete all the testing due to family or other emergencies. Therefore, the following samples were obtained:

N = 292 completed all cognitive and non-cognitive paper-and-pencil tests

N = 256 completed computer tests

N = 112-129 completed retest of paper-and-pencil tests (N varied across tests)

N = 113 completed retest of computer tests

N = 74 completed practice effects on computer tests

Table 2.9 shows the race and gender makeup for Fort Knox soldiers completing at least part of the Pilot Trial Battery. Table 2.10 shows the sample distribution by MOS code. The mean age of the participating soldiers was 21.9 years (SD = 3.1). The mean years in service, computed from date of enlistment in the Army, was 1.6 years (SD = 0.9).

Table 2.9

Race and Gender of Fort Knox Field Test Sample of the Pilot Trial Battery

Race	Frequency
White	156
Hispanic	24
Black	121
American Indian	2
Total	303

Sex	Frequency
Female	57
Male	246

Table 2.10

Military Occupational Specialties of Fort Knox Field Test Sample of the Pilot Trial Battery

MOS	N	MOS	N
05B	1	63N	5
11B	2	63T	3
11C	3	63W	3
12B	16	63Y	1
13B	14	64C	10
13E	1	67G	1
19D	19	71D	1
19E	29	71G	4
19K	10	71L	21
31J	2	71M	3
31S	2	71N	1
31V	3	72E	1
35E	1	73C	2
36C	1	75B	7
36K	2	75D	1
41C	1	75F	1
43M	1	76C	11
44B	1	76P	2
44E	2	76V	9
45B	1	76W	1
45G	1	76Y	38
45K	1	81E	1
45N	8	82C	1
45T	1	84B	1
51B	3	91B	3
51N	1	91E	2
52D	1	92B	1
55B	2	93F	1
57E	1	94B	2
62B	2	94F	1
62E	1	95B	15
63B	8	96B	1
63D	1		
63E	4		
63J	1		

Additional Field Testing

As noted previously, field tests were conducted at three sites. The sites and the basic purpose of the field test at each site were as follows:

Fort Knox. The full Pilot Trial Battery was administered here, as described above.

Fort Bragg. The non-cognitive Pilot Trial Battery measures, Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE), were administered to soldiers at Fort Bragg under several experimental conditions in order to estimate the extent to which scores on these instruments could be altered or "faked," when persons are instructed to do so. Information on procedures and sample is contained in Chapter 8.

Minneapolis Military Entrance Processing Station (MEPS). The non-cognitive measures were administered to a sample of soldiers as they were being processed into the Army in order to estimate how persons might alter their scores in an actual applicant setting. Information on procedures and sample is contained in Chapter 8.

Summary

The field test was completed in September 1984. Appendix G contains a copy of the Pilot Trial Battery as it was administered during the field tests.

The remaining chapters in this report describe the development of the Pilot Trial Battery measures, the analyses of the pilot test and field test data, and the revisions made to the battery based on those analyses.

CHAPTER 3

COGNITIVE PAPER-AND-PENCIL MEASURES: PILOT TESTING

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe,
Janis S. Houston, Norman G. Peterson, Teresa L. Russell,
and Mary Ann Hanson

GENERAL

This chapter deals with the cognitive paper-and-pencil measures developed for inclusion in the Pilot Trial Battery. As described in Chapter 1, the Task 2 research team, including contractor personnel, Army Research Institute monitors, and designated members of the Scientific Advisory Group, had previously evaluated and prioritized cognitive ability constructs or predictor categories according to their relevance and importance for predicting success in a variety of Army MOS (see Figure 1.5). These priority ratings were used to plan cognitive paper-and-pencil test development activities.

Before describing the development of the tests, we outline some issues and objectives germane to all the cognitive paper-and-pencil measures. Each cognitive predictor category is then discussed in turn.

Within each category, we provide a definition of the target cognitive ability. Next, for each test developed to measure the target ability, we outline the strategy followed; this included identifying (1) the target population or target MOS for which the measure is hypothesized to most effectively predict success, (2) published tests that served as markers for each new measure, (3) intended level of item difficulty, and (4) type of test (i.e., speed, power, or a combination). The test itself is then described and example items are provided. Results from the first two pilot test administrations or tryouts are reported to explain and document subsequent revisions. Finally, psychometric test data from the third pilot test, conducted at Fort Lewis, are discussed and the form of the test decided upon for field testing is described.

The last portion of this chapter presents a summary and analysis of the newly developed cognitive ability tests. This includes a discussion of test intercorrelations, results from a factor analysis of the intercorrelations, and results from subgroup analyses of test scores from the pilot test at Fort Lewis. Field testing of these measures is then described in Chapter 4.

Target Population

The population for which these tests have been developed is the same one to which the Army applies the ASVAB, that is, persons applying to enlist in the Army. This is, speaking very generally, a population made up of recent high school graduates, not entering college, from all geographic sections of the United States. Non-high-school graduates may be accepted into the Army, but present policy gives preference to high school graduates. For a number of reasons, Army applicants are probably not a truly random sample of all recent high school graduates, but for initial test development activities a highly refined specification of Army applicants was not necessary, and was not attempted.

Another point to be made about the target population is the fact that it was, practically speaking, inaccessible to us during our development process. We were constrained to use enlisted soldiers to try out the newly developed tests. Enlisted soldiers, of course, represent a restricted sample of the target population in that they all have passed enlistment standards; furthermore, almost all of the soldiers that we were able to use in our pilot tests had also passed Basic and Advanced Individual Training. Thus, the persons in our samples are presumably more qualified, more able, more persevering, and so forth, on the average, than are the persons in the target population.

The above discussion leads up to two major implications that served as general guidelines for our development and pilot testing activities:

(1) The tests to be developed will be applied to a population with a large range of abilities. Therefore, we should attempt to develop tests each of which have a broad range of item difficulties. Highly peaked tests, in the sense that all items would have difficulty levels near a certain value (e.g., .50, indicating that half the examinees would answer correctly), were not our goal.

(2) The soldiers upon whom the tests will be initially tried out are generally higher in ability than the target population. Therefore, the tests should be somewhat easier than they would be if we had access to an unrestricted sample of the target population in trying out the tests. With regard to this point, we point out the somewhat confusing nature of the technical term "difficulty level." This term is defined as the proportion of persons attempting an item who answer the item correctly. Thus, a high item difficulty level (say .90) means the item is relatively easy, whereas a low item difficulty level (say .10) means the item is relatively hard. When used in reference to an entire test, it is usually defined as the proportion of the total number of items that are answered correctly, on the average. Thus, a test difficulty level of .75 means that, on the average, persons taking the test answer 75% of the items correctly.

Power vs. Speed

The above discussion of the target population shows how we derived some general guidelines about the difficulty level of the tests and their items. Another decision to be made about each test was its placement on the power vs. speed continuum. This decision is, of course, linked to the test difficulty issue, since a relatively easy test can usually be made difficult simply by reducing the time allowed to take the test.

Very few tests used in practical testing situations are pure power tests, but quite a few are highly speeded tests. Most psychometricians would agree that a "pure" power test is a test administered in such a way that all persons taking the test are allowed enough time to attempt all items on the test, and that a "pure" speeded test is a test administered in such a way that no one taking the test has enough time to attempt all of the items. In practice, there appears to be a power/speed continuum, and most tests fall somewhere between the two extremes on this continuum. It also is the case that a power test usually contains items that not all persons will be able to answer correctly, even given unlimited time to complete the test, while a speeded test usually contains items that all or almost all persons could answer correctly, given enough time to attempt the items.

As a matter of practical definition for this developmental effort, we used an "80% completion" rule-of-thumb to define a power test. That is, if a test could be completed by 80 percent of all those taking the test, then we considered it a "power" test. Tests with completion rates lower than this were considered to have some "speededness" determining performance on the test.

The Pilot Trial Battery contains cognitive ability tests that may be considered power tests, and tests that may be categorized as highly speeded tests, using the above definition. It also contains tests that may be viewed as combinations of both power and speed. Each test is defined below as a power, speeded, or combination test according to the development strategy employed.

Reliability

A final issue related to evaluation of test construction procedures is test reliability. Several procedures are available to assess the reliability of a measure and each provides distinctive information about a test. Internal consistency estimates are used to assess homogeneity of test content; high values indicate that test items are measuring the same ability or abilities. Test-retest procedures are used to estimate the stability of test scores across time; high values indicate that the test yields the same or very similar scores for each subject over time.

Split-half reliability estimates were obtained for each paper-and-pencil test administered at the pilot test sites: Fort Carson, Fort Campbell, and Fort Lewis. For each tryout, each test was administered in two separately timed parts. Reliability estimates were obtained by correlating scores from the two parts, and the Spearman-Brown correction procedure was then used to estimate the reliability for the whole test. The separately timed, split-half reliability estimates, corrected by the Spearman-Brown procedure, are reported for each test. This estimate of reliability is appropriate for either speeded or power tests.

Further, we also report Hoyt internal consistency reliability estimates for each test. This method provides the average reliability across all possible split-test halves. We point out that this procedure is inappropriate for speeded tests because it overestimates the reliability, but in the interest of complete reporting the Hoyt reliability estimate has been calculated for all tests.

Predictor Categories

We turn now to the description of the tests, which are discussed within cognitive ability constructs. The four constructs treated in cognitive paper-and-pencil tests were spatial visualization, field independence, spatial orientation, and induction/figural reasoning.

SPATIAL VISUALIZATION

Spatial visualization involves the ability to mentally manipulate components of two- or three-dimensional figures into other arrangements. The process involves restructuring the components of an object and accurately discerning their appropriate appearance in new configurations. This construct includes several subcomponents, two of which are:

- Rotation - the ability to identify a two-dimensional figure when seen at different angular orientations within the picture plane. It also includes three-dimensional rotation or the ability to identify a three-dimensional object projected on a two-dimensional plane, when seen at different angular orientations either within the picture plane or about the axis in depth.
- Scanning - the ability to visually survey a complex field to find a particular configuration representing a pathway through the field.

Visualization constructs had been given a mean validity estimate of .21 across all criterion constructs by our expert panel.[1] The highest mean validity estimate for visualization measures was .25 for criterion clusters involving Technical Skills.

Currently, no ASVAB measures are designed specifically to measure spatial abilities. For this reason, spatial visualization received a priority rating of one (see Figure 1.5), and development of spatial ability measures was strongly emphasized. The visualization construct was divided into two areas: visualization/rotation and visualization/scanning. We developed two tests to tap abilities within each of these areas; these four tests are described below.

Spatial Visualization - Rotation

The rotation component of spatial visualization requires the ability to mentally restructure or manipulate parts of a two- or three-dimensional figure. We developed two tests of this ability, Assembling Objects and Object Rotation. The former involves three-dimensional figures, and the latter involves two-dimensional objects.

Assembling Objects Test

Development Strategy. Predictive validity estimates provided by expert raters suggest that measures of the visualization/rotation construct would be effective predictors of success in MOS that involve mechanical operations (e.g., inspect and troubleshoot mechanical systems, inspect and troubleshoot electrical systems), construction (e.g., construct wooden buildings, construct masonry structures), and drawing or using maps. Thus,

---

[1] This panel was the group of 35 personnel psychologists who estimated the relationships between a set of ability constructs and a set of Army criterion constructs. See Chapter 1 of this report, also Wing, Peterson, and Hoffman (1984).

88

the Assembling Objects test was designed to yield information about the potential for success in MOS involving mechanical or construction activities.

Published tests identified as markers[2] for Assembling Objects include the Employee Aptitude Survey (EAS-5) Space Visualization and the Flanagan Industrial Test (FIT) Assembly. EAS-5 requires examinees to count three-dimensional objects depicted in two-dimensional space, whereas the FIT Assembly involves mentally piecing together objects that are cut apart or disassembled. The FIT Assembly was selected as the more appropriate marker for our purposes because it has both visualization and rotation components for mechanical or construction activities. Thus, we designed the Assembling Objects Test to assess the ability to visualize how an object will look when its parts are put together correctly.

Multiple-choice test items were constructed to tap this ability at several difficulty levels ranging from very easy items to more difficult items. It was determined that this measure would combine power and speed components, with speed receiving greater emphasis.

Test Development. In the original form of the Assembling Objects Test, subjects were asked to complete 30 items within a 16-minute time limit. Each item presented subjects with components or parts of an object. The task was to select from among four alternatives the one object that depicted the components or parts put together correctly. Two item types were included in the test; examples of each are shown in Figure 3.1.

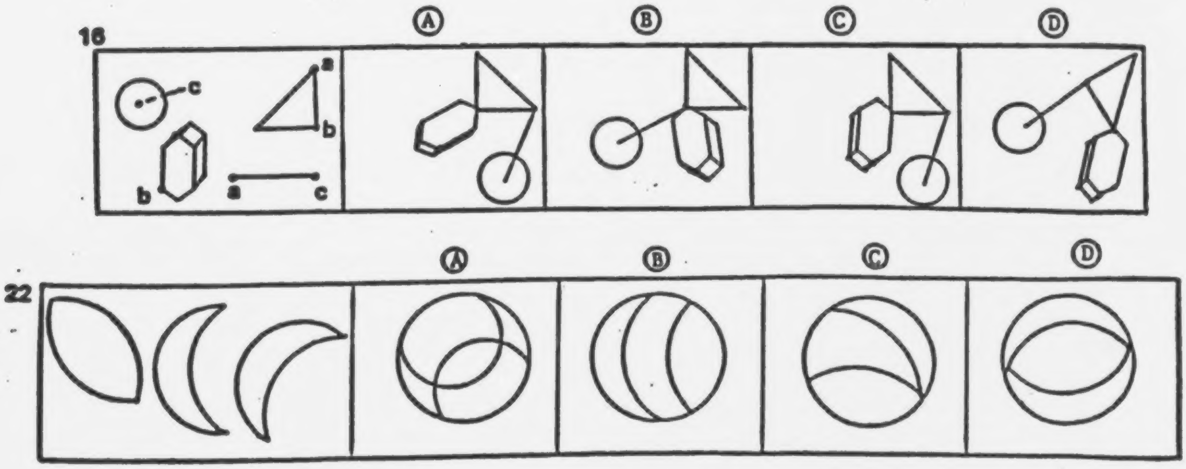


Figure 3.1. Sample Items from Assembling Objects Test.

---

[2] As mentioned in Chapter 2, marker tests were published tests that were judged to measure the predictor categories on constructs for which we were developing tests. Some of these marker tests were actually administered during pilot testing, others were not, but they were all studied to assist in developing the new tests.

The first tryout, conducted at Fort Carson, indicated that the test may have suffered from ceiling effects. That is, nearly all recruits in this sample (N = 36) completed the test; the mean score was 24.2 (SD = 5.05). Further, item difficulty levels were somewhat higher than intended (mean = .80, SD = .12, median = .83); that is, the proportion of examinees obtaining high scores was greater than expected.

Therefore, ten new, more difficult items, five for each item type, were constructed and added to the test to reduce the likelihood of ceiling effects. The 16-minute time limit was retained for the second tryout, at Fort Campbell. Nearly all subjects (N = 56) completed the test (mean items completed = 37.3, SD = 4.75); the mean score was 26.3 (SD = 8.34). Item difficulty levels were lower for the revised test (mean = .68, SD = .15, median = .72). Inspection of these results indicated that the test possessed acceptable psychometric qualities, so no further changes were made in preparation for the Fort Lewis pilot test.

Pilot Test Results. Fort Lewis results for the Assembling Objects Test are shown in Table 3.1. The test contains 40 items with a 16-minute time limit; individual test scores were computed using the total number correct. The mean number of items completed was 37.6, with a range of 18 to 40. Corresponding values for number correct (or test score) were 28.1 and 7-40.

Parts 1 and 2 correlate .65 with each other. Reliabilities are estimated at .79 by split-half methods (Spearman-Brown corrected), and .89 with Hoyt's estimate of reliability.

For the total test, item difficulties (see Figure 3.2) range from .31 to .92 with a mean of .70. We also computed the correlation of scores on each item (0 = incorrect, 1 = correct) with total scores (the number of items answered correctly). This index, usually called the item-total correlation, measures the degree to which each item is measuring the same ability or abilities as the other items on the test. The higher the value of this index, the "better" the item. Values of .25 or better are usually considered acceptable, though lower values are not necessarily unacceptable. Item-total correlations for Assembling Objects range from .18 to .60 with a mean of .44 (SD = 9.99).

Correlations between scores on this measure and scores on other Pilot Trial Battery paper-and-pencil measures are reported at the end of this chapter. It is important, however, to note the correlations between this test and its marker tests. Both marker tests were administered in the Fort Carson tryout and the FIT Assembly was also used at Fort Campbell. Results from Fort Carson indicate that scores on the Assembling Objects Test correlate .74 with scores on EAS-5 and .76 with scores on FIT Assembly. Results from Fort Campbell indicate that this test correlates .64 with FIT Assembly. This last value represents a better estimate of the relationship between Assembling Objects and the FIT Assembly marker, because of the revisions made to Assembling Objects following the first tryout at Fort Carson. Given the sample sizes involved and the goals for the Assembling Objects Test, the .64 correlation was encouraging.

Modifications for the Fort Knox Field Test. In preparation for the Fort Knox administration, some Assembling Objects items were redrawn to

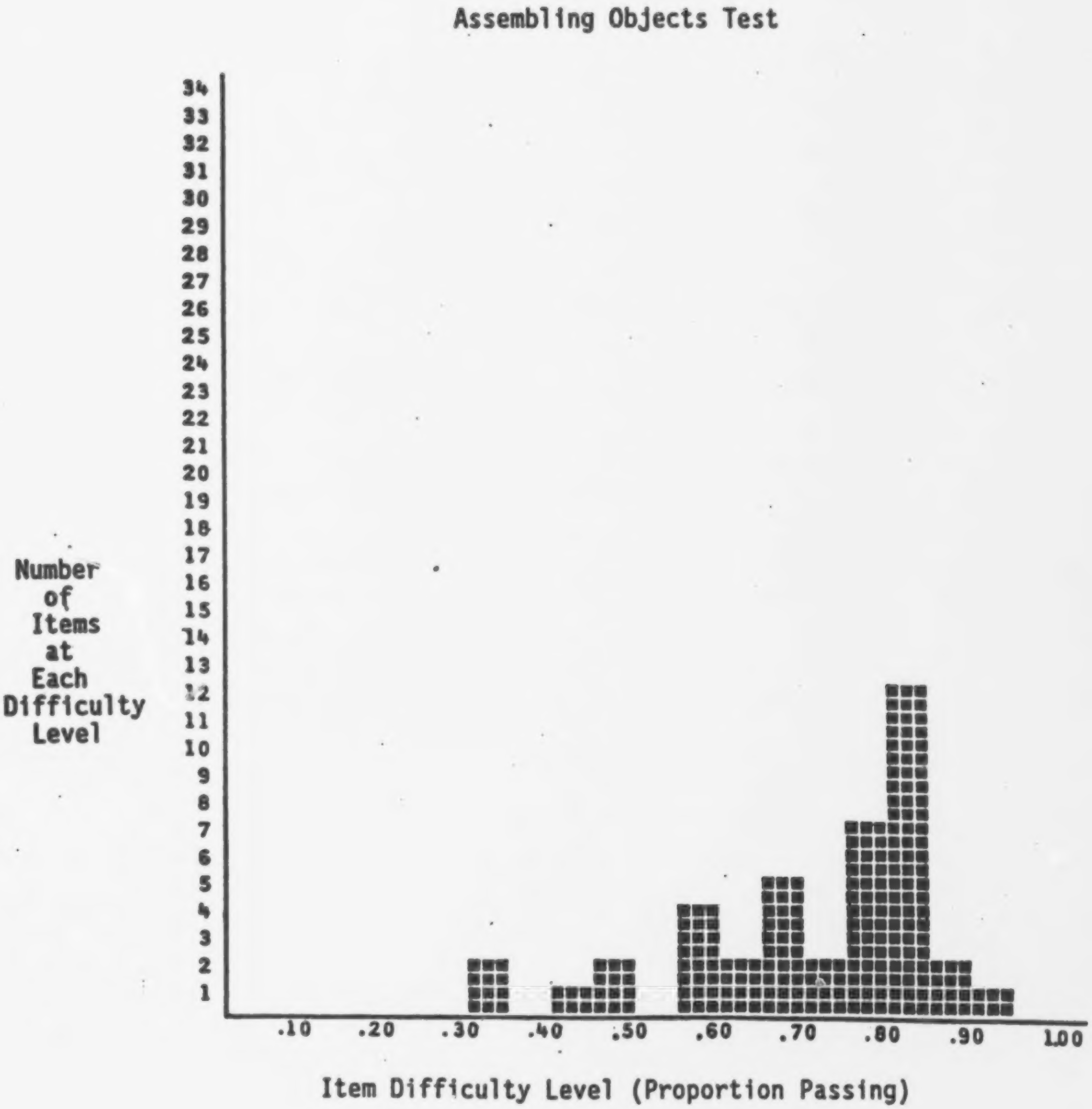
Table 3.1

Pilot Test Results from Fort Lewis: Assembling Objects Test

	Total	Part 1	Part 2
Number of Items	40	20	20
Time Allowed (minutes)	16 min.	8 min.	8 min.
Number of Subjects	118	118	118
Number of Items Completed			
Mean	37.58	18.23	19.36
Standard Deviation	3.83	2.59	2.12
Range	18-40	10-20	6-20
Last Item Completed by 80% of the Sample	N/A	16	20
Percentage of Subjects Completing All Items	48%	56%	80%
Number of Items Correct			
Mean	28.14	13.86	14.29
Standard Deviation	7.51	4.18	4.09
Range	7-40	3-20	3-20
Total-Part Intercorrelations			
Total	**	.91	.90
Part 1		**	.65
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .79

Hoyt Internal Consistency = .89



Mean = .70

SD = .16

Range = .31 - .92

NOTE: Number of items in the test = 40.

Figure 3.2. Distribution of item difficulty levels: Assembling Objects Test.

clarify the figures. The item response format was modified to a form that could be used for machine scoring (i.e., the subject was instructed to fill in a circle for the correct answer). This change was made in all of the tests being prepared for field test administration.

Object Rotation Test

Development Strategy. Object Rotation is the second test developed to measure spatial visualization/rotation. This measure is also expected to predict success in MOS involving mechanical operations, construction activities, and drawing or using maps.

Published tests serving as markers for this measure include Educational Testing Services (ETS) Card Rotations, Thurstone's Flags Test, and the Shephard-Metzler Mental Rotations. Each of these measures requires the subject to compare a test object with a standard object to determine whether the two represent the same figure with one simply turned or rotated or whether the two represent different figures. The first two measures, ETS Card Rotations and Thurstone's Flags, involve visualizing two-dimensional rotation of an object, whereas the Mental Rotations test requires visualizing three-dimensional objects depicted in two-dimensional space.

Object Rotation Test items were constructed to reflect a limited range of item difficulty levels ranging from very easy to moderately easy. These items, on the average, were designed to be easier than those appearing in the Assembling Objects Test. Further, we planned to construct a test that contains more items and has a shorter time limit than the Assembling Objects Test. Thus, the plan for Object Rotation was to develop a test that falls more toward the speeded end of the power-speed continuum.

Test Development. As initially developed, the Object Rotation Test contained 60 items with a 7-minute time limit. The subject's task was to examine a test object and determine whether the figure represented in each item is the same as the test object, only rotated, or is not the same as the test object (e.g., is flipped over). For each test object there are five test items, each requiring a response of "same" or "not same." Sample test items are shown in Figure 3.3.

The Fort Carson tryout indicated that this test suffered from ceiling effects. Subjects (N = 38), on the average, completed 59.3 (SD = 2.60) of the 60 items and obtained a mean score of 55.6 (SD = 6.06). Item difficulty levels averaged .92 (SD = .05). Consequently, we decided to add 30 new items to the test and to increase the time limit to 9 minutes for the second tryout at Fort Campbell.

In the second tryout, subjects, on the average, completed 87.6 (SD = 7.96) of the 90 items and obtained a mean score of 77.0 (SD = 12.1). The time limit was reduced to 8 minutes for the Fort Lewis administration, in order to obtain a more highly speeded test.

Pilot Test Results. Detailed results from the Fort Lewis pilot test are shown in Table 3.2. As reported in the table, completion rates were fairly high (mean = 82.6), with a range of 48 to 90. Test scores, computed by the total number correct, range from 36 to 90 with a mean of 73.4.

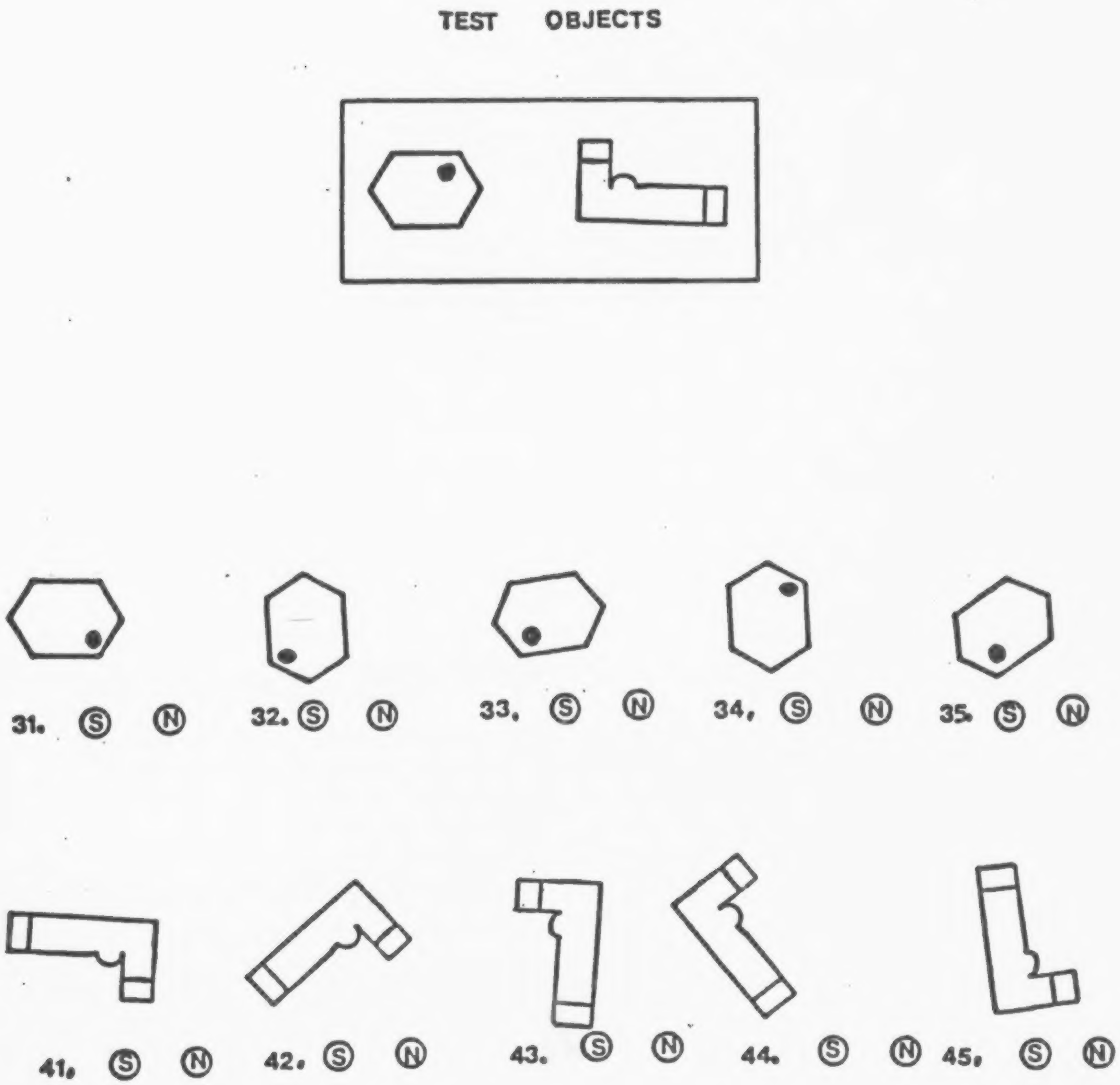

TEST OBJECTS
31. Ⓢ Ⓝ
32. Ⓢ Ⓝ
33. Ⓢ Ⓝ
34. Ⓢ Ⓝ
35. Ⓢ Ⓝ
41. Ⓢ Ⓝ
42. Ⓢ Ⓝ
43. Ⓢ Ⓝ
44. Ⓢ Ⓝ
45. Ⓢ Ⓝ

Figure 3.3. Sample Test items from Object Rotations Test.

Table 3.2

Pilot Test Results from Fort Lewis: Object Rotation Test

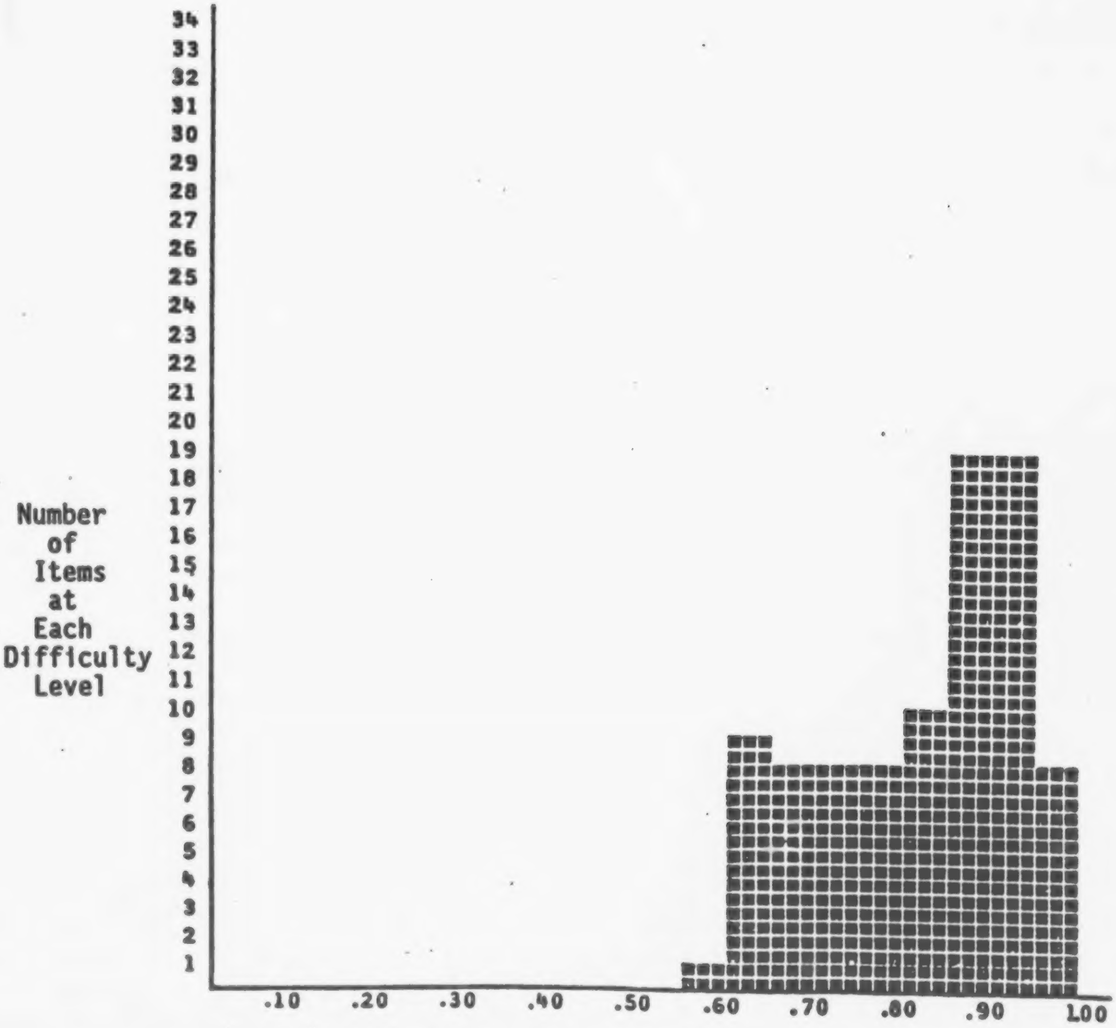
	Total	Part 1	Part 2
Number of Items	90	45	45
Time Allowed (minutes)	8	4	4
Number of Subjects	118	118	118
Number of Items Completed			
Mean	82.64	40.52	42.12
Standard Deviation	10.79	6.73	5.56
Range	48-90	21-45	18-45
Last Item Completed by 80% of the Sample	N/A	35	40
Percentage of Subjects Completing All Items	52%	60%	67%
Number of Items Correct			
Mean	73.36	36.64	36.72
Standard Deviation	15.40	8.69	7.77
Range	36-90	13-45	7-45
Total-Part Intercorrelations			
Total	**	.94	.93
Part 1		**	.75
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .86

Hoyt Internal Consistency = .96

a(c)

Object Rotation Test



Item Difficulty Level (Proportion Passing)

Mean = .82

SD = .11

Range = .59 - .98

NOTE: Number of items in the test = 90.

Figure 3.4. Distribution of item difficulty levels: Object Rotation Test.

Item difficulty levels (see Figure 3.4) range from .59 to .98 with a mean of .81. Item-total correlations averaged .44 (SD = .17), ranging from .09 to .79. Parts 1 and 2 correlated .75 with each other. The split-half reliability estimate, corrected for test length, is .86 while the Hoyt estimate is .96.

The marker test for Object Rotation, Mental Rotations, was administered at two of the three pilot test sites. Data collected at the Fort Carson tryout indicate that the two measures correlate .60 (N = 30); data from the Fort Lewis administration indicate the two correlate .56 (N = 118). This was viewed as an acceptable level of relationship.

Modifications for the Fort Knox Field Test. Results from the Fort Lewis pilot test indicated that the Object Rotation Test items possessed desirable psychometric properties. Number of items completed, item difficulties, and item-total correlations were nearly all acceptable. However, the time limit was decreased to 7 1/2 minutes to make the test more speeded and avoid a possible ceiling effect. Also, as noted earlier, the response format was modified to one that could be used for machine scoring.

Spatial Visualization - Scanning

A second component of spatial visualization ability which was emphasized in predictor development is spatial scanning. Spatial scanning tasks require the subject to visually survey a complex field and find a pathway through it, utilizing a particular configuration. The Path Test and the Maze Test were developed to measure this component of spatial visualization.

Path Test

Development Strategy. Validity estimates provided by the expert rating panel suggested that a measure of visualization/scanning would be most effective in predicting success for Army MOS involving electrical or electronic operations (e.g., troubleshooting electrical systems, inspecting and troubleshooting electronic systems), using maps in the field (e.g., planning placement of tactical positions), and controlling air traffic.

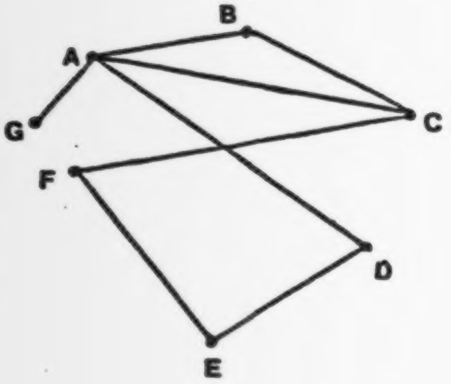
Published tests serving as markers for construction of the Path Test include Educational Testing Service's Map Planning and Choosing a Path. In these measures, examinees are provided with a map or diagram. The task is to follow a given set of rules or directions to proceed through the pathway or to locate an object on the map.

Results from the Preliminary Battery research with the marker tests, ETS Map Planning and ETS Choosing a Path, indicated that both tests are highly speeded and were very difficult for the target sample (Hough, Dunnette, Wing, Houston, & Peterson, 1984). For example, 80 percent of the subjects (N = 1,843 Army recruits) completed only 16 of the 40 items contained in the Map Planning test. The mean score for this group was 18.1 (SD = 16.5). For Choosing a Path, 80 percent of the subjects completed only six of the 16 items. This group obtained a mean score of 4.96 (SD = 3.35).

These data suggested that the Path Test should contain items somewhat less difficult than the ETS tests or provide more time for completion of items at a similar difficulty level. Consequently, Path Test items were constructed to yield difficulty levels for the target population ranging from very easy to somewhat difficult, and the test time was established to place more emphasis on speed than on power.

Test Development. The Path Test requires subjects to determine the best path or route between two points. Subjects are presented with a map of airline routes or flight paths. Figure 3.5 shows a flight path with four sample items. The subject's task is to find the "best" path--that is, the path between two points that requires the fewest stops. Each lettered dot is a city that counts as one stop; the beginning and ending cities (dots) do not count as stops.

In its original form, the Path Test contained 35 items with a 9-minute time limit. Subjects were asked to record the numbers of stops for each item in the corresponding blank space. (The response format appearing in Figure 3.5 is from the final version of the Path Test.) The first version contained three maps or airline routes with 13, 9, and 13 items, respectively.



The route from:	Number of Stops:
1. A to F	● ② ③ ④ ⑤
2. G to E	① ② ③ ④ ⑤
3. C to D	① ② ③ ④ ⑤
4. G to F	① ② ③ ④ ⑤

Figure 3.5. Sample items from Path Test.

Results from the first tryout, conducted at Fort Carson, revealed that the test was too easy. Virtually all of the subjects completed the test (mean = 34.1, SD = 2.51, N = 21) and the mean score was 29.9 (SD = 4.08). Item difficulty levels ranged from .48 to 1.00 with a mean of .85 (SD = .12).

To reduce the potential for ceiling effects, an additional map or flight path with 13 items was added to the test. Also, four very easy items (i.e., difficulty levels ranging from .90 to 1.00) were deleted, resulting in 44 items on the revised test. The 9-minute time limit was retained. In the second tryout subjects completed an average of 40.7 items (SD = 5.07) and obtained a mean score of 32.6 (SD = 7.00). Item difficulty levels ranged from .55 to .96 with a mean of .80. Those results indicated that the changes had largely achieved the goal of making the test more difficult.

To prepare for the pilot test conducted at Fort Lewis, the test response format was revised to allow subjects to circle the number of stops (i.e., 1-5) to avoid having to process written-in responses. In addition, the time limit was reduced from 9 minutes to 8 minutes to increase the speededness of the test.

Pilot Test Results. Path Test results obtained from the Fort Lewis tryout are reported in Table 3.3. Subjects, on the average, completed 35.3 of the 44 items, with a range of 0 to 44. Test scores, computed by the total number correct, ranged from 0 to 44 with a mean of 28.3.

Item difficulty levels (see Figure 3.6) ranged from .20 to .91 with a mean of .64). Item-total correlations averaged .47 (SD = 11) with a range of .25 to .69. Parts 1 and 2 correlate .70. The split-half reliability estimate, corrected for test length, is .82. The Hoyt internal consistency value is .92. These results indicated that the test is generally in excellent shape.

Both marker tests were administered at the first tryout, and the ETS Map Planning Test was also administered at the Fort Campbell and Fort Lewis tryouts. Data from the first tryout indicate that the original Path Test correlates .34 with ETS Choosing a Path and -.01 with ETS Map Planning. The reader is reminded that results from Fort Carson are based on a very small sample size (N = 19) and that the Path Test was modified greatly following this tryout. Data from the final two tryouts indicate that the Path Test and Map Planning correlate .62 (N = 54) and .48 (N = 118), respectively. Although these values are not as high as marker test correlations for some of the other new tests, this was expected. Recall that the marker tests were known to be too difficult for the typical Army sample and we set out to make the new tests easier than the marker tests.

Modifications for the Fort Knox Field Test. The Path Test remained unchanged for the field test except for the modification in response format.

Maze Test

Development Strategy. The Maze Test represents the second measure constructed to assess spatial visualization/scanning. As with the Path Test, the expert panel of judges indicated that this measure would be most effective in predicting success for MOS involving electrical and electronic operations, using maps in the field, and controlling air traffic.

The development strategy for this test mirrors that of the Path Test--markers for the Maze Test again included ETS Map Planning and ETS Choosing

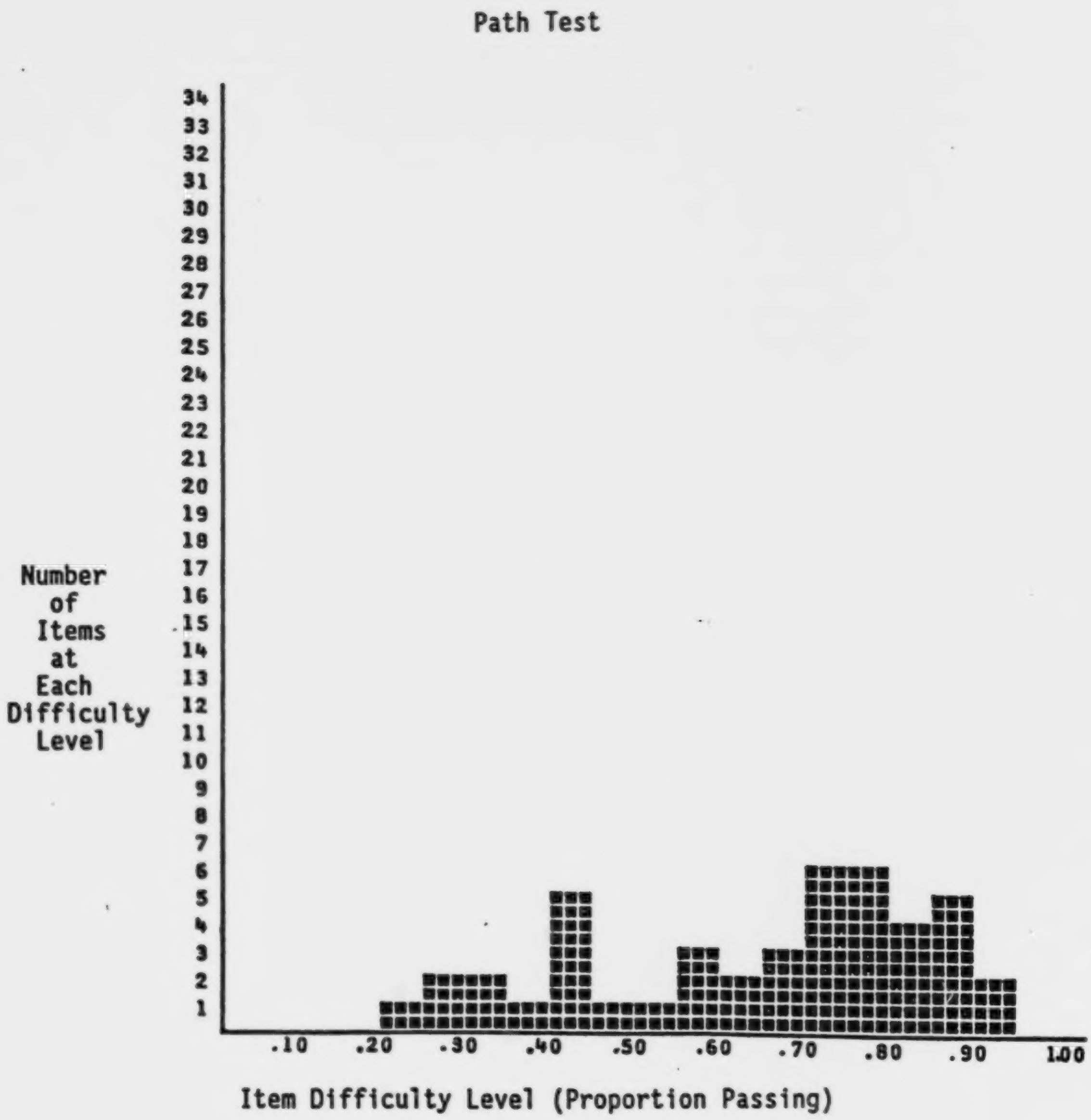
Table 3.3

Pilot Test Results from Fort Lewis: Path Test

	Total	Part 1	Part 2
Number of Items	44	22	22
Time Allowed (minutes)	8	4	4
Number of Subjects	116	116	116
Number of Items Completed			
Mean	35.33	16.63	18.70
Standard Deviation	8.27	4.58	4.25
Range	0-44	0-22	0-22
Last Item Completed by 80% of the Sample	N/A	13	15
Percentage of Subjects Completing All Items	19%	23%	42%
Number of Items Correct			
Mean	28.28	13.41	14.87
Standard Deviation	9.08	4.93	4.91
Range	0-44	0-22	0-22
Total-Part Intercorrelations			
Total	**	.92	.92
Part 1		**	.70
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .82

Hoyt Internal Consistency = .92



Mean = .64

SD = .20

Range = .20 - .91

NOTE: Number of items in the test = 44.

Figure 3.6. Distribution of item difficulty levels: Path Test.

a Path. As with the Path Test, this test was designed to include items geared more toward the ability level of the Project A target population than populations for the two marker tests, that is, somewhat easier items were appropriate for the Maze Test.

However, the Maze Test differs from the Path Test in several ways. The task required in the Maze Test involves finding the one pathway that allows exit from a maze. Items for the Maze Test were constructed to be much easier under nonspeeded conditions than in the Path Test, and greater emphasis was placed on speed. The Maze Test, then, was designed to measure visualization/scanning ability under highly speeded conditions.

Test Development. For the first tryout the Maze Test contained 24 rectangular mazes. Each included four entrance points labeled A, B, C, and D, and three exit points indicated by an asterisk (*). The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points. A 9-minute time limit was established for this test.

Results from the first tryout, at Fort Carson, indicate that the original version of the Maze Test suffered from ceiling effects. Subjects completed on average 23.3 (SD = 1.79) of the 24 items and obtained a mean score of 22.6 (SD = 2.75).

To increase test score variance, the test was modified in two ways. First, an additional exit was added to each test maze; Figure 3.7 shows a sample item from the original test and the same item modified for the Fort Campbell tryout. Second, the time limit was reduced from 9 to 8 minutes.

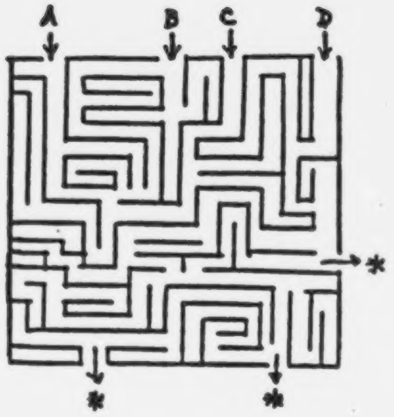
At the second tryout, completion rates were again high (mean = 22.5, SD = 2.49, N = 56). Consequently, for the third tryout, the time limit for completing the 24 maze items was dropped to 6 minutes.

Pilot Test Results. Results from the Fort Lewis administration are reported in Table 3.4. These data indicate that the reduced time limit produced a drop in the completion rate for the Fort Lewis sample (mean = 20.7. Test scores, computed by the total number correct, ranged from 8 to 24 with a mean of 19.3.

Item difficulty levels (see Figure 3.8) range from .41 to .98 with a mean of .80. Item-total correlations average .48 (SD = .22) with a range of -.04 to .80. Parts 1 and 2 correlate .64 with each other. The split-half reliability estimate corrected for test length is .78 and the Hoyt reliability estimate for this test is .88. Taken as a whole, these results indicate that the test is in good shape.

One or both of the marker tests, ETS Choosing a Path and ETS Map Planning, were administered at the three pilot test sites. Results from Fort Carson indicate that the Maze Test correlates .24 (N = 29) with Choosing a Path, and .36 (N = 30) with Map Planning. These values must be viewed with caution because of the small sample size and because of modifications made to the Maze Test following this tryout. Map Planning was also administered at the Fort Campbell and Fort Lewis tryouts. Data collected

FORT CARSON



FORT CAMPBELL

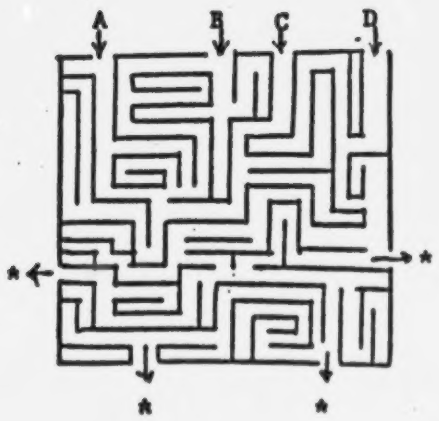


Figure 3.7. Sample items for the Maze Test.

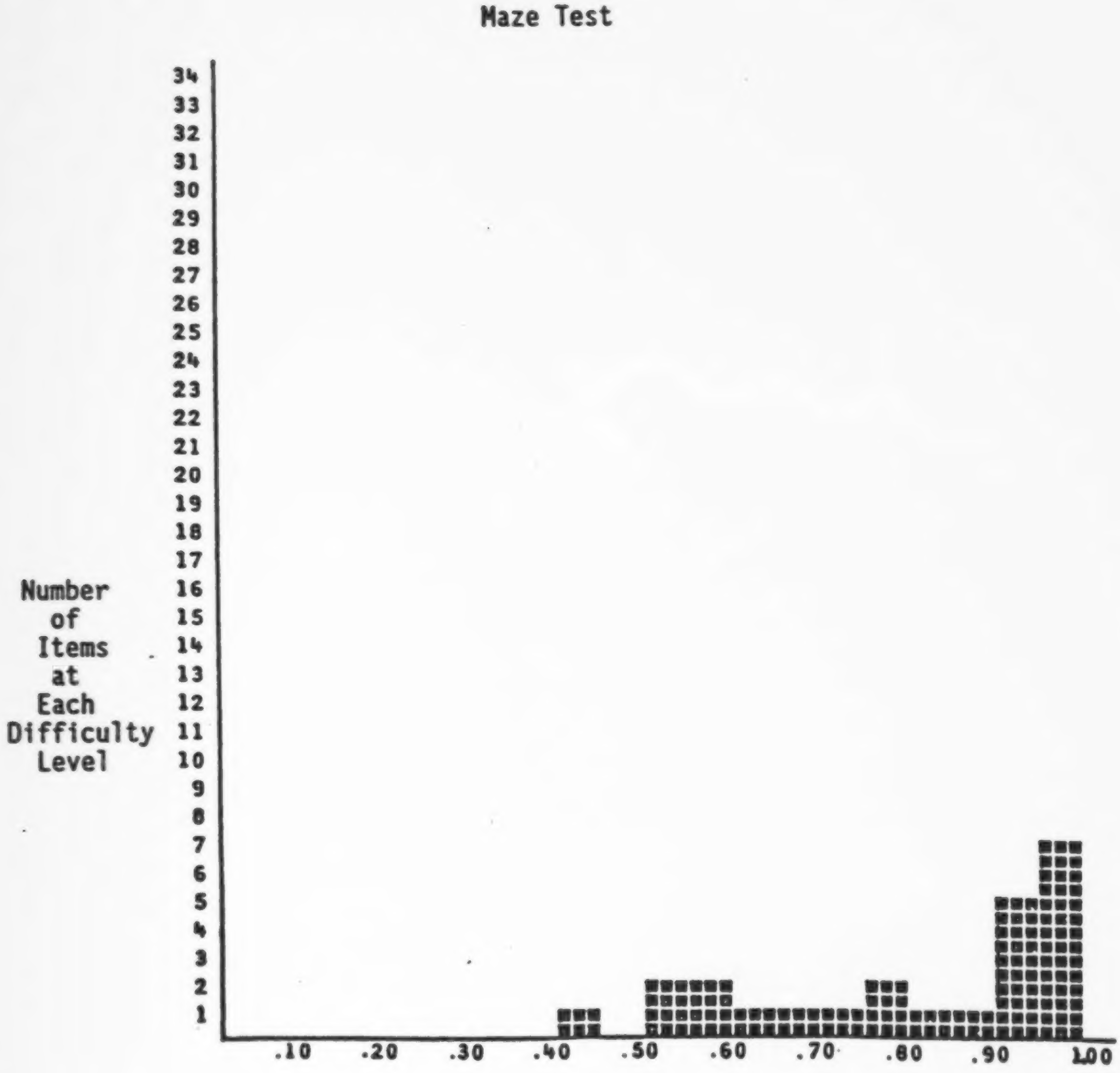
Table 3.4

Pilot Test Results from Fort Lewis: Maze Test

	Total	Part 1	Part 2
Number of Items	24	12	12
Time Allowed (minutes)	6	3	3
Number of Subjects	118	118	118
Number of Items Completed			
Mean	20.65	10.44	10.21
Standard Deviation	3.88	2.18	2.19
Range	9-24	3-12	4-12
Last Item Completed by 80% of the Sample	N/A	9	8
Percentage of Subjects Completing All Items	38%	57%	50%
Number of Items Correct			
Mean	19.30	9.95	9.35
Standard Deviation	4.35	2.48	2.32
Range	8-24	2-12	4-12
Total-Part Intercorrelations			
Total	**	.91	.90
Part 1		**	.64
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .78

Hoyt Internal Consistency = .88



Item Difficulty Level (Proportion Passing)

Mean = .80

SD = .18

Range = .41 - .98

NOTE: Number of items in the test = 24.

Figure 3.8. Distribution of item difficulty levels: Maze Test.

at these posts indicate that it correlates .45 (N = 55) and .63 (N = 118), respectively, with the revised Maze Test. This last correlation was viewed as acceptable.

Modifications for the Fort Knox Field Test. Results from the last pilot test administration showed that the Maze Test could be slightly more speeded. The percentage of subjects completing this test was higher than for the Path Test (38% for the Maze Test, and 19% for the Path). Therefore, the time limit was reduced from 6 minutes to 5 1/2 minutes for the Fort Knox field test.

FIELD INDEPENDENCE

This construct involves the ability to find a simple form when it is hidden in a complex pattern. Given a visual percept or configuration, field independence refers to the ability to hold the percept or configuration in mind so as to disembed it from other well-defined perceptual material.

This construct received a mean validity estimate of .30 from the panel of expert judges, with the highest estimate of .37 appearing for MOS that involve detecting and identifying targets. Field Independence received a priority rating of two for inclusion in the battery.

Shapes Test

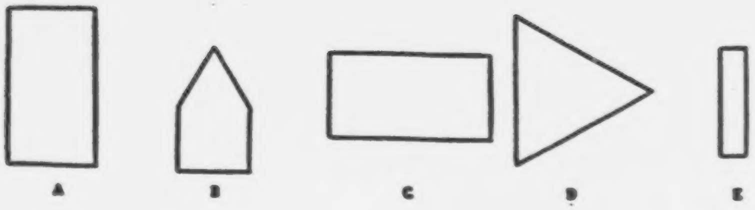
Development Strategy. According to the expert panel of judges, a measure of field independence most effectively predicts success for MOS that involve detecting and identifying targets, using maps in the field, planning placement of tactical position, controlling air traffic, and troubleshooting operating systems such as mechanical, electrical, fluid, and electronic systems.

The marker test for the Shapes Test is the Educational Testing Service's Hidden Figures Test, a measure included in the Preliminary Battery (Hough, et al., 1984). In this test, subjects are asked to find one of five simple figures located in a more complex pattern. Initial analyses of the Preliminary Battery indicated that for the target population of first-term enlisted soldiers, the Hidden Figures Test suffers from limited test score variance, and possibly floor effects. For example, the initial data indicate that 80 percent of the sample completed fewer than 4 of the 16 test items. The mean test score was, therefore, very low (mean = 5.16, SD = 3.35).

Our strategy for constructing the Shapes Test, then, was to use a task similar to that in the Hidden Figures Test while ensuring that the difficulty level of test items was geared more toward the Project A target population. Further, we decided to include more types of items than appear in the Hidden Figures Test and to construct items that reflect varying difficulty levels ranging from easy to moderately difficult. We wanted the test to be speeded, but not nearly so much so as the ETS Hidden Figures Test.

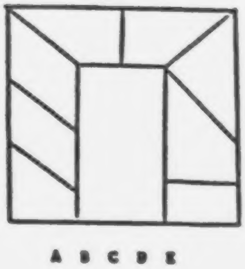
Test Development. At the top of each test page are five simple shapes; below these shapes are six complex figures. Subjects are instructed to examine the simple shapes and then to find the one simple shape located in each complex figure. (See Figure 3.9.)

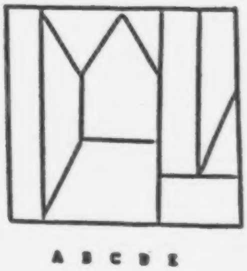
In the first tryout, at Fort Carson, the Shapes Test contained 54 items with a 16-minute time limit. Results from this tryout indicated that most subjects were able to complete the entire test (e.g., mean completed = 53.4, SD = 1.53), and most subjects obtained very high scores (mean score = 49.3, SD = 4.17). Item difficulty levels also suggested that this test was very easy and suffered from ceiling effects (mean item difficulty level = .91, SD = .13, median = .97).



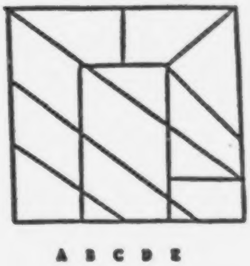
Complex Figures

Ft. Carson





Ft. Campbell



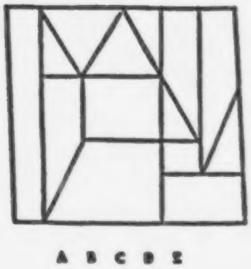


Figure 3.9. Sample items from the Shapes Test.

To prepare for the Fort Campbell tryout, nearly all test items were modified to increase item difficulty levels. Examples of item modifications are provided in Figure 3.9. As is shown, by adding a few lines to each complex pattern, the test items administered at Fort Campbell tryout were made more difficult than the items administered at Fort Carson.

Results from Fort Campbell indicate that test item modifications were successful. Subjects, on the average, completed 43.5 (SD = 8.79) of the 54 items within the 16-minute time limit, and obtained a mean score of 30.7 (SD = 23.5, and median difficulty level = .67).

This test was modified only slightly for the Fort Lewis administration. For example, a few complex figures inadvertently contained more than one simple figure. (This was revealed in the item analyses.) These items were revised to ensure that no more than one simple figure could be located in each complex figure. The Shapes Test administered to the Fort Lewis sample contained 54 items with a 6-minute time limit.

Pilot Test Results. Table 3.5 contains Fort Lewis results from the Shapes Test. Mean number completed is 42.4. The mean number correct for this sample is 29.3 with a range of 12 to 51, indicating that the measure does not suffer from ceiling effects.

Item difficulty levels (see Figure 3.10) range from .10 to .97 with a mean of 54.2 (SD = 24.55). (See Figure 3.10.) Item-total correlations range from .07 to .57 with a mean of .39 (SD = .13). Reliability estimates indicate that Parts 1 and 2 correlate .69; with the Spearman-Brown correction, this value is .82. The Hoyt reliability estimate for this test is .89. As a whole, these results show the test to be in good shape.

The marker test, ETS Hidden Figures Test, was administered at the first two tryouts. Results from Fort Carson indicate that the original version of the Shapes Test correlated .35 with the Hidden Figures Test (N = 29). Data from Fort Campbell indicate that the revised Shapes Test correlates .50 with its marker (N = 56). Although a bit lower than desirable, this was not unexpected because of the planned differences in difficulties of the two tests.

Modifications for the Fort Knox Field Test. The Shapes Test needed only minor revisions for the field test. For example, item-total correlations for a few items indicated that more than one shape could still be located in a complex figure test item, so these figures were modified.

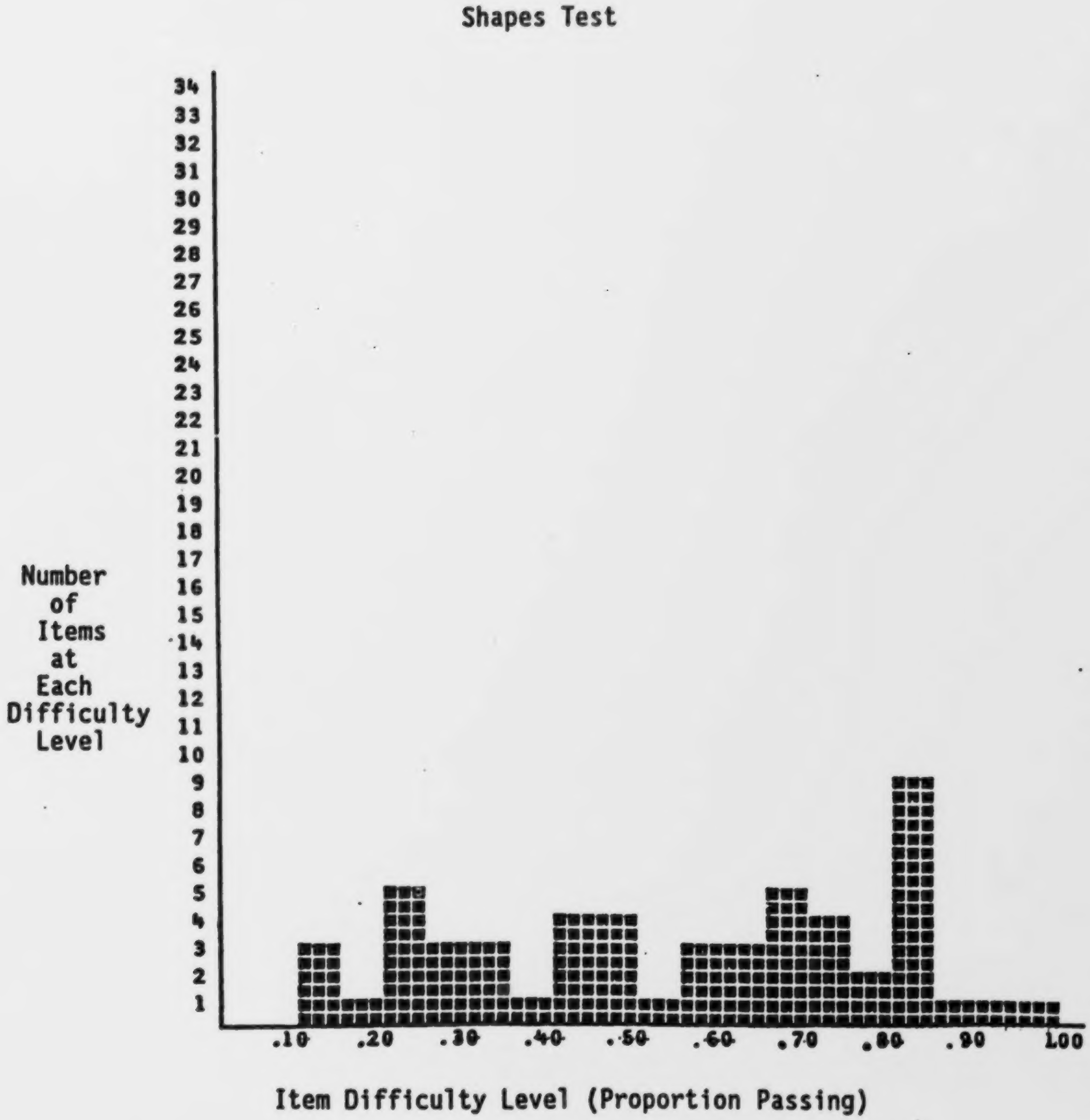
Table 3.5

Pilot Test Results from Fort Lewis: Shapes Test

	Total	Part 1	Part 2
Number of Items	54	27	27
Time Allowed (minutes)	16	8	8
Number of Subjects	118	118	118
Number of Items Completed			
Mean	42.42	20.78	21.64
Standard Deviation	9.29	5.14	5.05
Range	17-54	8-27	8-27
Last Item Completed by 80% of the Sample	N/A	16	17
Percentage of Subjects Completing All Items	12%	24%	23%
Number of Items Correct			
Mean	29.28	14.49	14.79
Standard Deviation	9.14	5.03	4.92
Range	12-51	5-26	4-25
Total-Part Intercorrelations			
Total	**	.92	.92
Part 1		**	.69
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .82

Hoyt Internal Consistency = .89



Mean = .54

SD = .25

Range = .10 - .97

NOTE: Number of items in the test = 54.

Figure 3.10. Distribution of item difficulty levels: Shapes Test.

SPATIAL ORIENTATION

This construct involves the ability to maintain one's bearings with respect to points on a compass and to maintain appreciation of one's location relative to landmarks in the environment.

This particular construct was not included in the list of predictor constructs evaluated by the expert panel. The rationale for developing measures of spatial orientation for inclusion in the Pilot Trial Battery is described below.

Conceptualization and measurement of this ability construct first appeared during World War II, when researchers for the Army Air Force (AAF) Aviation Psychology Program explored a variety of constructs to aid in the selection of air crew personnel. Spatial orientation measures were designed to predict success in air crew positions that required familiarity with points on a compass, the ability to apprehend directions quickly and accurately, and the ability to remain directionally oriented in spite of sudden and frequent changes in direction. Results from the AAF Program indicated that measures of spatial orientation were useful in selecting pilots and navigators (Guilford & Lacey, 1947).[3]

During the second year of Project A, several Task 2 personnel from PDRI had the opportunity to observe recruits performing on the job[4]. These job observations included soldiers from a variety of MOS, such as administrative specialists, cannon crewmen, armor crewmen, radio and teletype operators, light wheel vehicle/power generator equipment mechanics, infantrymen, military police, and MANPADS personnel. Information collected during these job observations suggested that some MOS involve critical job requirements of maintaining directional orientation and establishing location using features or landmarks in the environment. For example, armor or tank crewmen when performing in the field must be able to reorient themselves quickly as the tank turret turns or rotates; MANPADS personnel need to establish their location in the field, relative to the location of friendly and enemy troops, using features or landmarks in the environment.

Information obtained from these job observations was reported, in part, at the March 1984 Task 2 IPR. Participants in this meeting agreed that measures of spatial orientation would be useful in predicting performance in Army MOS that require orientation abilities if a soldier is to be successful on the job. Three measures were developed for this construct.

Orientation Test 1

Development Strategy. As reported above, information collected during job observations suggested that a measure of spatial orientation would be most effective in predicting success for MOS that include such critical job requirements as identifying tactical positions, determining location of friendly and enemy troops, and using features or landmarks in the environment to establish and maintain one's bearings.

---

[3] Dr. Lloyd Humphreys, of the Scientific Advisory Group for Project A, particularly emphasized the usefulness of this construct to us.

[4] Dr. Jay Uhlaner, also of SAG, originally suggested that job observation sessions would be especially helpful at this stage of the research, which indeed proved to be the case.

Paper-and-pencil measures that tap this ability were developed by researchers in the U.S. Army Air Force's Aviation Psychology Program. Direction Orientation Form B (CP515B) served as the marker for Orientation Test 1. The strategy for developing Orientation 1 involved generating items that duplicated the task in the Army Air Force's test. Each item contained six circles. The first, the standard compass or "given" circle, indicates the direction of North and usually is rotated out of the conventional position. The remaining circles are test compasses that also have directions marked on them.

For this test, item construction was limited to one of seven possible directions: South, East, West, Southwest, Northwest, Southeast, and Northeast. Thus, item difficulty levels were not expected to vary greatly. (Off-quadrant directional items such as Northwest or Southeast were, however, viewed as more difficult than South, East, or West directional items.) Our plan for this test was to ask subjects to complete numerous compass directional items within a short period of time. Orientation 1, then, was designed as a highly speeded test of spatial orientation.

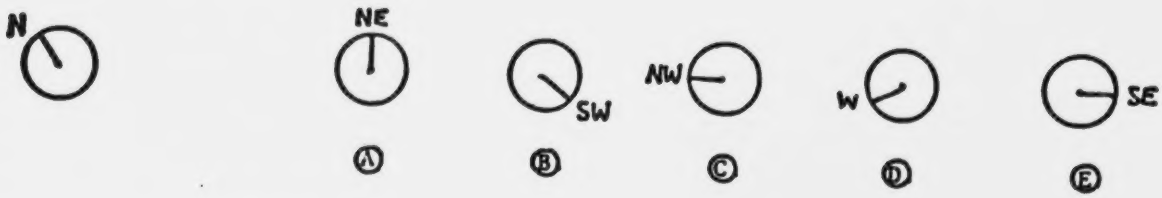
Test Development. In its original form, each test item presented subjects with six circles. The first, the Given Circle, indicated the compass direction for North. For most items, North was rotated out of its conventional position (i.e., the top of the circle did not necessarily represent North). Compass directions also appeared on the remaining five circles. The subject's task was to determine, for each circle, whether or not the direction indicated was correctly positioned by comparing it to the direction of North in the Given Circle. (See Example 1 in Figure 3.11.)

When administered to the Fort Carson sample, this test contained 20 item sets requiring 100 responses (i.e., for every item, compass directions on five circles must be evaluated). Subjects were given 8 minutes to complete the test. Test scores were determined by the total number correct; the maximum possible was 100.

Results from this first tryout showed that nearly all subjects completed the items within the time allotted (mean completed was 18.6 out of the 20 sets of items); they obtained a mean score of 82.7 (SD = 17.9). Item difficulty levels indicate that most items were moderately easy (mean = 82.7, SD = 11.1).

Thus, for the Fort Campbell tryout, we attempted to create more difficult items by modifying directional information provided in the Given Circle. That is, rather than indicating the direction for North, compass directions for South, East, or West were provided. These directions were also rotated out of conventional compass position. (See Example 2, Figure 3.11.)

EXAMPLE 1


N
NE
SW
NW
W
SE
Ⓐ
Ⓑ
Ⓒ
Ⓓ
Ⓔ

EXAMPLE 2

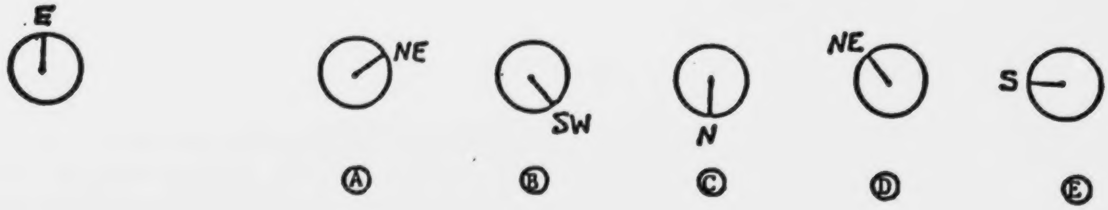


Figure 3.11. Sample items from Orientation Test 1.

Orientation Test 1, as administered at the Fort Campbell tryout, contained 30 item sets (150 items). It was administered in three separately timed parts. Parts One and Two included the original test items, whereas Part Three included the new (non-North) items. This last part of the test was preceded by additional test instructions that informed subjects about the change in Given Circle directions. Subjects were given 3 minutes to complete each part, for a total of 9 minutes.

Results from this second tryout indicate that for the total test, subjects completed 23.5 of the 30 item sets (or 117.10 items) and obtained a mean score of 100.8 (SD = 24.0). Scores on Part Three yielded lower correlations with Parts One and Two (both are .44); Parts One and Two correlated .87. From this information we reasoned that the new items were assessing additional information about subjects' abilities to maintain orientation.

We then mixed item sets from Part Three with item sets from Parts One and Two to create a test with 30 item sets (150 items) for the Fort Lewis tryout. The time limit was increased to a total of 10 minutes, and test instructions were modified to explain that items vary throughout the test with respect to information provided in the Given Circle. Again, test score was determined by the number of items correct (maximum score is 150).

Pilot Test Results. Results from the Fort Lewis pilot test are reported in Table 3.6. Completion rates/for the total test indicated that, on the average, subjects attempted 25 of the 30 item sets (or 125.7 of 150 items) and obtained a mean score of 117.9 (SD = 24.2).

Item difficulty levels (see Figure 3.12) ranged from .21 to .97 with a mean of .79. Item-total correlations are at acceptable levels (mean = .43, SD = .14). The correlation between Parts One and Two is .86. Reliability estimates are as follows: Split-half Spearman-Brown corrected = .92, Hoyt = .97. These results indicate that the test was performing as intended.

No marker tests for this construct were included in any of the three pilot test administrations. However, two other new measures of spatial orientation (Orientation 2 and Orientation 3) were developed for the Pilot Trial Battery and correlations between Orientation 1 and these other new tests were obtained. (These new tests are described below.) From the Fort Carson data, Orientation 1 correlated .40 with Orientation 2 (N = 30) and .66 with Orientation 3 (N = 25). Results from Fort Campbell indicate that Orientation 1 correlated .45 with Orientation 2 and .72 with Orientation 3 (N = 56). Finally, for the Fort Lewis sample, these same measures correlated .53 and .68, respectively (N = 118). These results were viewed as indicating that Orientation 1 was tapping the appropriate constructs, but was not redundant with the other new tests.

Modifications for the Fort Knox Field Test. Very few changes were made on this test; for example, one item was "cleaned up" to avoid confusion about the compass direction provided on the Given Circle. The field test version of Orientation Test 1 contained 30 item sets (150 items) with a 10-minute time limit.

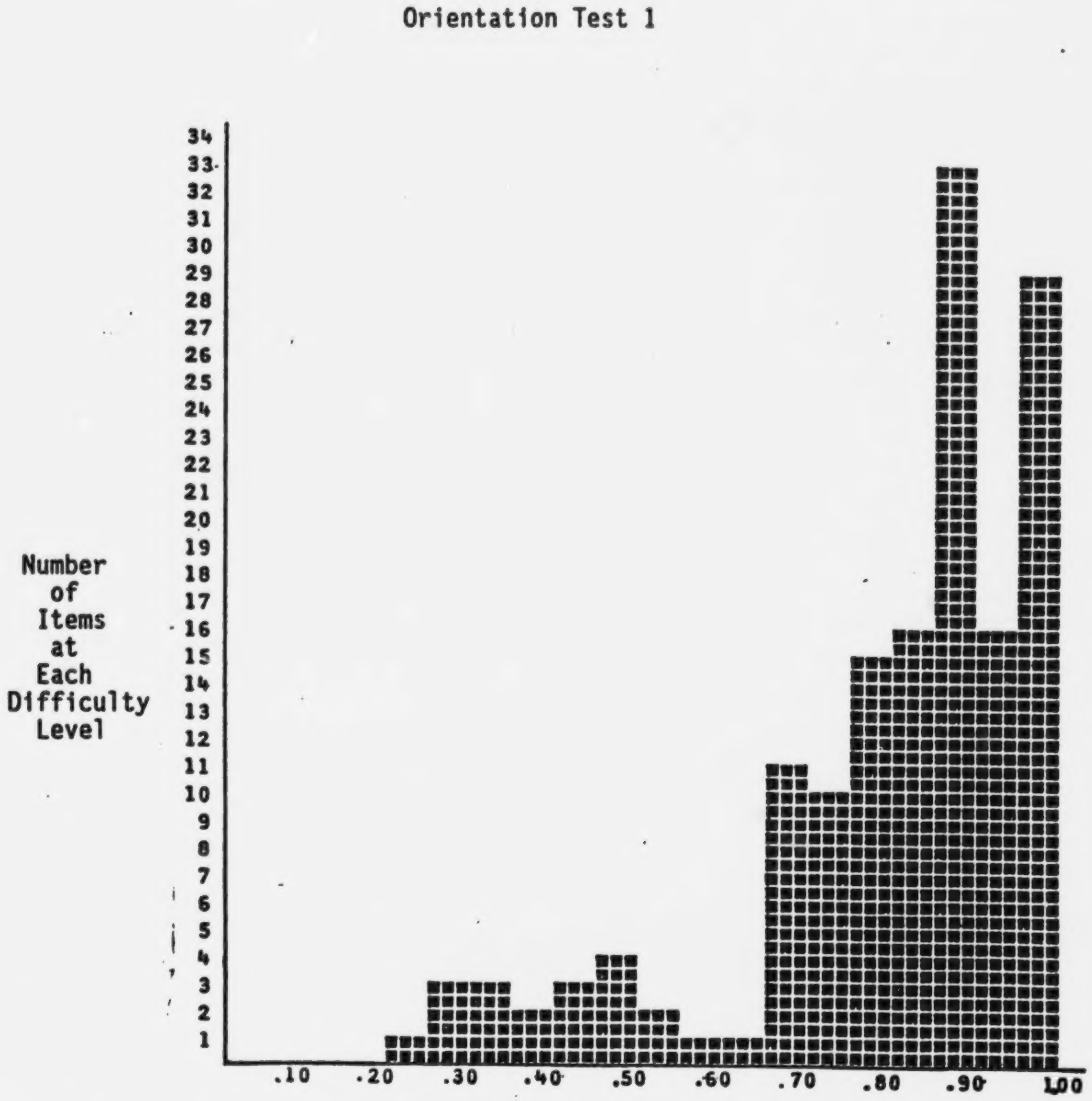
Table 3.6

Pilot Test Results from Fort Lewis: Orientation Test 1

	Total	Part 1	Part 2
Number of Items	30(150)	15(75)	15(75)
Time Allowed (minutes)	10	5	5
Number of Subjects	118	118	118
Number of Items Completed			
Mean	25.14(125.7)	11.75(58.75)	13.39(66.95)
Standard Deviation	4.88	2.96	2.35
Range	12-30(60-150)	5-15(25-75)	5-15(25-75)
Last Item Completed by 80% of the Sample	N/A	9(45)	12(60)
Percentage of Subjects Completing All Items	31%	32%	55%
Number of Items Correct			
Mean	117.86	56.50	61.36
Standard Deviation	24.16	12.28	12.80
Range	46-150	25-75	21-75
Total-Part Intercorrelations			
Total	**	.96	.96
Part 1		**	.89
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .92

Hoyt Internal Consistency = .97



Mean = .79

SD = .20

Range = .21 - .98

NOTE: Number of items in the test = 150.

Figure 3.12. Distribution of item difficulty levels: Orientation Test 1.

Orientation Test 2

Development Strategy. The second measure of spatial orientation was also designed to tap abilities that might predict success for MOS that involve maintaining appreciation of one's location relative to landmarks in the environment or in spite of frequent changes in direction. Orientation Test 2 is a relatively new approach to assessing spatial orientation abilities.

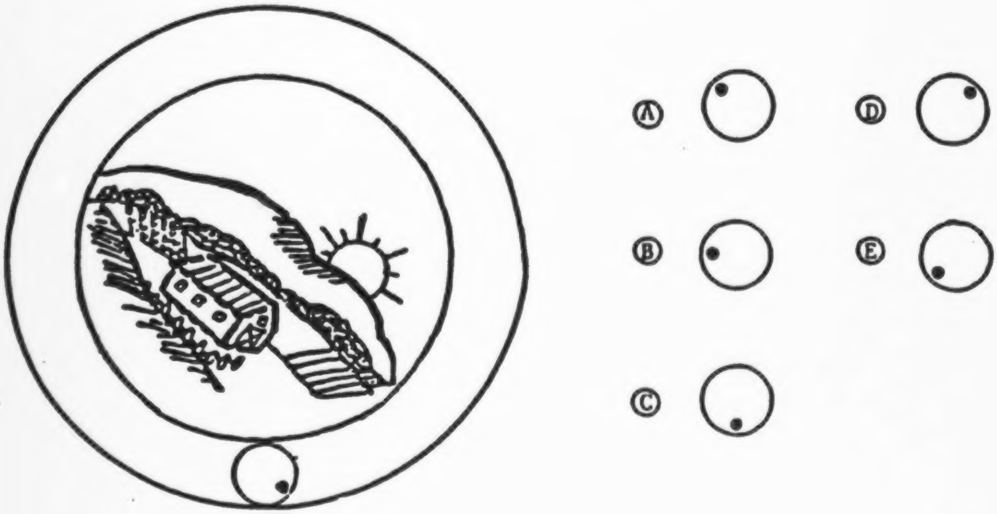
Although no particular test served as its model, it is similar to a measure designed by Army Air Force researchers to select pilots, navigators, and bombardiers (Directional Orientation: CP5150). Items in the AAF test consist of two aerial photographs of the same landscape. On the first photograph, a compass is indicated. The second photograph is rotated relative to the first photograph and contains an arrow, again indicating direction. Subjects must determine in which direction the arrow in the second picture is pointed, based on the compass direction given in the first photograph and the degree of rotation of the second photograph. Thus, the AAF test measures the ability to maintain one's perspective with regard to the directional relationships of several objects (e.g., the first aerial photograph) when the objects have been rotated (e.g., the second aerial photograph).

The task we designed for Orientation Test 2 asks subjects to mentally rotate objects and then to visualize how components or parts of those objects will appear after the object is rotated. Item difficulty levels were varied by altering the degree of rotation required to correctly complete each part of the task. Because of the complexity of the task, Orientation 2 was initially viewed as a power test of spatial orientation.

Test Development. For Orientation Test 2, we chose to design a task involving common objects. Each item contains a picture within a circular or rectangular frame. At the bottom of the frame is a circle with a dot inside it. The picture or scene is not in an upright position. The task is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture; after doing so, one must then determine where the dot will appear in the circle. (See Figure 3.13 for sample items.) For the Fort Carson tryout, this test contained 20 items with an 8-minute time limit.

Results from this administration indicate that the time limit was sufficient (mean number completed = 19.9, SD = 4.55). Item difficulty levels were somewhat lower than desired (mean = .52, SD = .16). Item-total correlations were, however, impressive (mean = .48, SD = .10). The only potential problem with this measure involved the test instructions as some subjects required additional instructions to understand what was going on. Therefore, for the Fort Campbell tryout, test instructions were modified to clarify the task.

Data collected at Fort Campbell provide very similar information about this test. For example, nearly all subjects completed this test (mean = 19.7, SD = .71). Item-total correlations were again impressive (mean = .46, SD = .13). The mean score and item difficulty levels indicated that the test was more difficult for this group than for the Fort Carson sample (mean score = 8.61, SD = 4.49; mean item difficulty = .43, SD = .11).



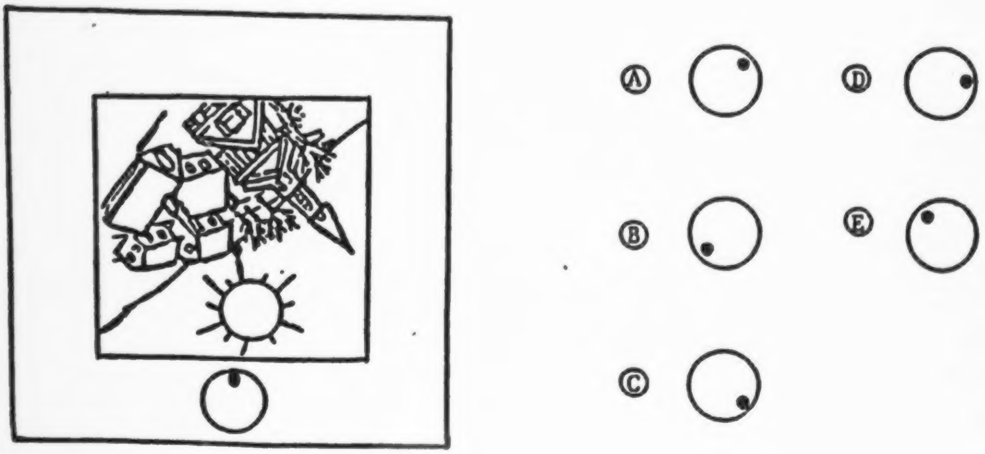


Figure 3.13. Sample items from Orientation Test 2.

Because of these item difficulty levels, we decided to add four new test items, constructed using item difficulty information obtained for the Fort Campbell sample. That is, items were examined to identify what appeared to make them more or less difficult, and new, easier items were written using this information. Primarily, this involved constructing items so that rotations of 90, 180, or 270 degrees were correct.

Orientation Test 2, as administered to the Fort Lewis sample, contained 24 items. A 10-minute time limit was established to correspond to the increase in the number of items. Test scores on this measure are determined by the total number correct.

Pilot Test Results. Table 3.7 contains the results from the Fort Lewis test. These data indicate that Orientation 2 is a power test (mean number completed = 23.7, SD = 1.04). Subjects obtained a mean score of 11.5 (SD = 6.20).

Item difficulty levels (see Figure 3.14) ranged from .19 to .71 with a mean of .48. This represents a slight increase from the Fort Campbell tryout, indicating the test was somewhat easier. Item-total correlations remained high, ranging from .22 to .74 with a mean of .53. Scores from Parts 1 and 2 correlate .80. Correcting this value for test length yields a split-half reliability estimate of .89. The Hoyt internal consistency value is also .89. Thus, this test has excellent reliability and distributional properties and met its goal of being a power test.

As noted above, no marker tests for this test were administered in any of the three tryouts. Correlations with the other newly developed measures of spatial orientation were obtained at each tryout. Data from Fort Carson indicate that Orientation 2 correlates .40 with Orientation 1 (N = 29) and .42 with Orientation 3. Results from Fort Campbell indicate that these same measures correlate .45 and .54 (N = 56). Finally, the Fort Lewis data indicate the measures correlate .53 and .65 (N = 118). These correlations were viewed as about right, that is Orientation Test 2 did correlate moderately with other Orientation tests but not so high as to be redundant.

Modifications for the Fort Knox Field Test. For the Fort Knox administration, this measure was unchanged except for the usual modification of the response format.

Orientation Test 3

Development Strategy. This test was also designed to measure spatial orientation. As with the other two measures of this construct, Orientation Test 3 is expected to be useful in predicting success for MOS that involve establishing and maintaining one's bearing using features or landmarks in the environment.

Orientation Test 3 was modeled after another spatial orientation test, Compass Directions, developed by researchers in the Army Air Force's Aviation Psychology Program. The AAF measure was designed to assess the ability to reorient oneself to a particular ground pattern quickly and accurately when compass directions are shifted about. Orientation 3 was designed to assess the same ability, using a similar test format.

120

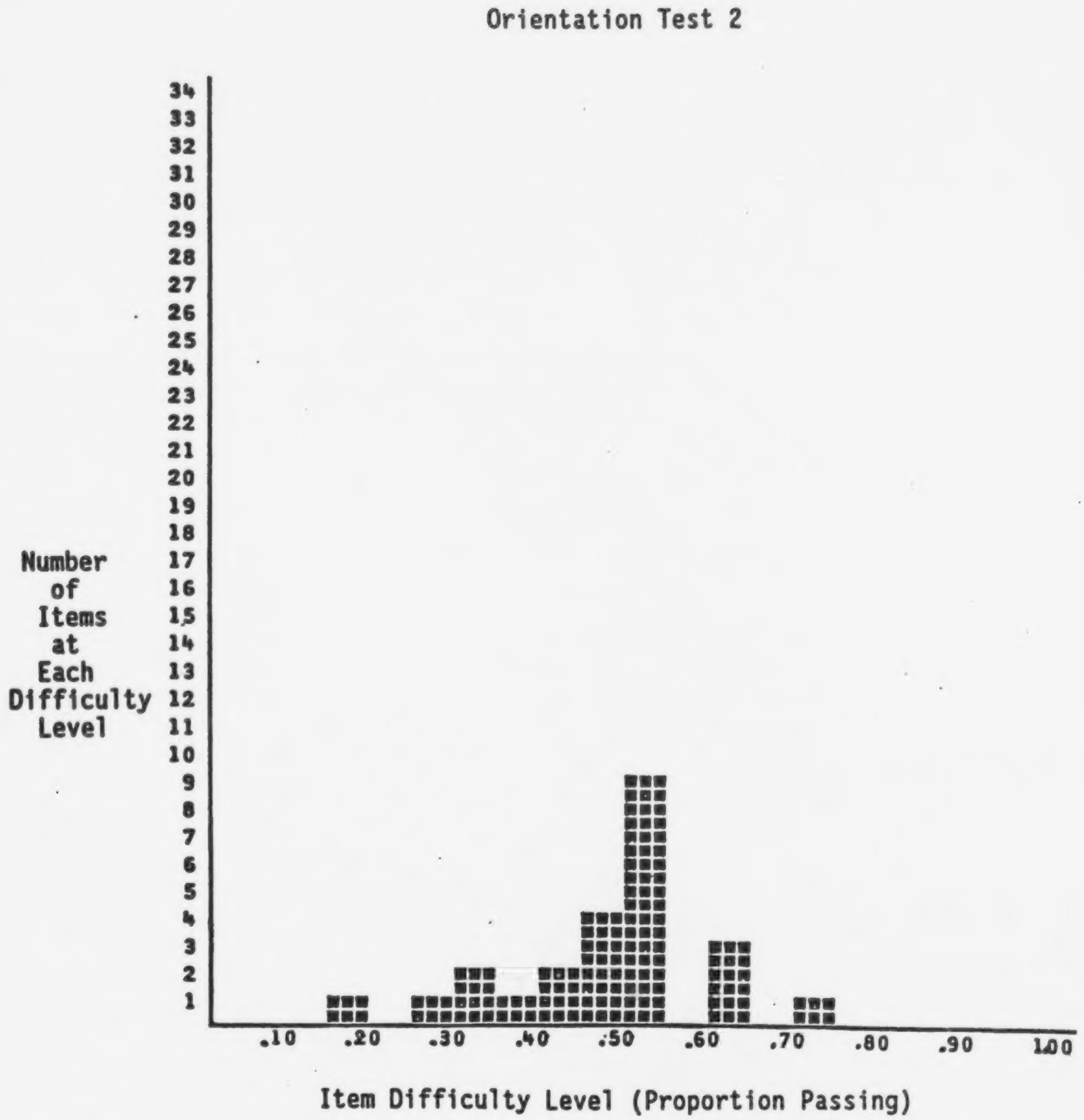
Table 3.7

Pilot Test Results from Fort Lewis: Orientation Test 2

	Total	Part 1	Part 2
Number of Items	24	12	12
Time Allowed (minutes)	10	5	5
Number of Subjects	118	118	118
Number of Items Completed			
Mean	23.73	11.85	11.88
Standard Deviation	1.04	.71	.45
Range	16-24	6-12	9-12
Last Item Completed by 80% of the Sample	N/A	12	12
Percentage of Subjects Completing All Items	90%	93%	92%
Number of Items Correct			
Mean	11.53	5.37	6.16
Standard Deviation	6.20	3.25	3.28
Range	3-24	0-12	0-12
Total-Part Intercorrelations			
Total	**	.95	.95
Part 1		**	.80
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .89

Hoyt Internal Consistency = .89



Mean = .48

SD = .12

Range = .19 - .71

NOTE: Number of items in the test = 24.

Figure 3.14. Distribution of item difficulty levels: Orientation Test 2.

Items for Orientation 3 were constructed to yield varying difficulty levels from moderately easy to moderately difficult. This test was designed to place somewhat more emphasis on speed than on power.

Test Development. In its original form, Orientation 3 presented subjects with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake, and so on. Within each item, subjects are provided with compass directions by information on the direction of one landmark with respect to another, such as "the forest is north of the campsite." Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark. Figure 3.15 contains one test map and two sample items. Note that for each item, new or different compass directions are given.

For the Fort Carson tryout, the test contained two maps with 10 questions about each map, for a total of 20 items. Subjects were given 12 minutes to complete the test. Results from this first tryout revealed very few problems with the test (e.g., test instructions were clear, the time was sufficient, no floor nor ceiling effects appeared). Thus, this measure remained unchanged for the Fort Campbell pilot test.

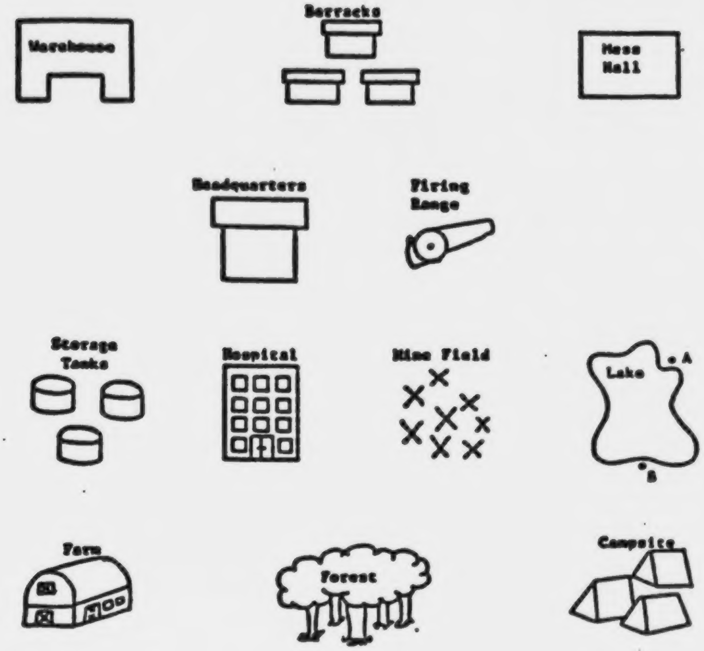
Results from the second tryout yielded similar information (e.g., no ceiling nor floor effects, acceptable completion rates). These data, however, indicated that for a few items, two responses might be correct due to a lack of precision in drawing the two maps. Accordingly, landmarks on each map were repositioned to ensure that one and only one correct answer existed for each item. In addition, one item was rewritten to make its wording uniform with other test items. When administered to the Fort Lewis sample, Orientation 3 contained 20 test items with a 12-minute time limit. Test scores are determined by the total number correct.

Pilot Test Results. Results from the Fort Lewis administration are reported in Table 3.8. On the average, subjects completed 18 items. The mean score of 8.7 indicates that subjects correctly answered about one-half of the items attempted.

Item difficulty levels (see Figure 3.16) range from .24 to .63 with a mean of .44. Item-total correlations range from .48 to .72 with a mean of .59 (SD = .07). Part 1 and Part 2 correlate .79. The split-half reliability estimate corrected for test length is .88, while the Hoyt internal consistency estimate is .90. These results indicate that the test is highly reliable, had acceptable distributional properties, and was appropriately speeded.

Data from Fort Carson indicate that Orientation Test 3 correlates .66 with Orientation 1 (N = 29) and .42 with Orientation 2 (N = 31). Values for these same measures administered at Fort Campbell are .72 and .54 (N = 56). Data from Fort Lewis indicate that these measures correlate .68 and .65 (N = 118). As with the other two Orientation tests, these results were viewed as acceptable.

Modifications for the Fort Knox Field Test. This test was unchanged for the Fort Knox field test except for the response format modifications.


Warehouse
Barracks
Mess Hall
Headquarters
Firing Range
Storage Tanks
Hospital
Mine Field
Lake
A
B
Farm
Forest
Campsite

1. The forest is due west of the barracks. You are at headquarters. Which direction must you travel in order to reach the firing range?

   1. N 2. NE 3. E 4. SE 5. S 6. SW 7. W 8. NW

2. The firing range is southwest of the hospital. You are at the farm. Which direction must you travel in order to reach the campsite?

   1. N 2. NE 3. E 4. SE 5. S 6. SW 7. W 8. NW

Figure 3.15. Sample items from Orientation Test 3.

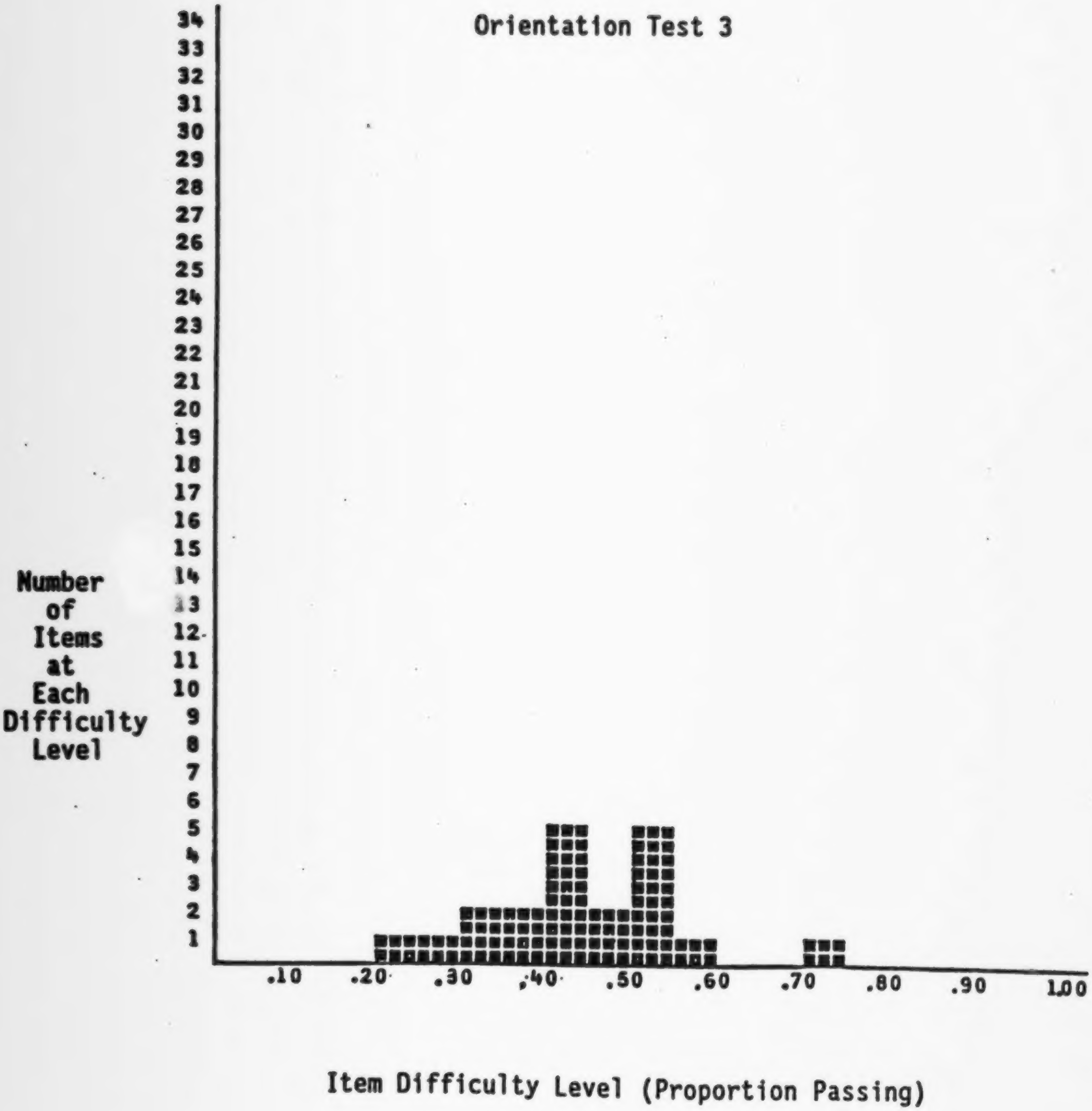
Table 3.8

Pilot Test Results from Fort Lewis: Orientation Test 3

	Total	Part 1	Part 2
Number of Items	20	10	10
Time Allowed (minutes)	12	6	6
Number of Subjects	118	118	118
Number of Items Completed			
Mean	18.12	8.82	9.30
Standard Deviation	2.68	1.76	1.26
Range	8-20	2-10	4-10
Last Item Completed by 80% of the Sample	N/A	7	9
Percentage of Subjects Completing All Items	42%	52%	67%
Number of Items Correct			
Mean	8.71	3.99	4.72
Standard Deviation	5.78	2.93	3.19
Range	0-20	0-10	0-10
Total-Part Intercorrelations			
Total	**	.94	.95
Part 1		**	.79
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .88

Hoyt Internal Consistency = .90



Mean = .44

SD = .10

Range = .24 - .63

NOTE: Number of items in the test = 20.

Figure 3.16. Distribution of item difficulty levels: Orientation Test 3.

INDUCTION - FIGURAL REASONING

This construct involves the ability to generate hypotheses about principles governing relationships among several objects.

Example measures of induction include the Employee Aptitude Survey Test 6 - Numerical Reasoning (EAS-6), Educational Testing Service's Figure Classification, Differential Aptitude Test (DAT) Abstract Reasoning, Science Research Associates (SRA) Word Grouping, and Raven's Progressive Matrices. These paper-and-pencil measures present subjects with a series of objects such as figures, numbers, or words. To complete the task, subjects must first determine the rule governing the relationship among the objects and then apply the rule to identify the next object in the series.

The panel of expert judges indicated that a measure of inductive reasoning would be useful for predicting success in numerous Army MOS. Specifically, for figural reasoning these judges estimated the mean validity at .25. The Army's current selection and classification system measures reasoning ability using word problems, but lacks a general measure of hypothesis generation and application. Two measures of reasoning were developed.

Reasoning Test 1

Development Strategy. According to the panel of experts, a measure of figural reasoning should effectively predict success in a wide variety of MOS, especially those that involve troubleshooting, inspecting and repairing operations systems, analyzing intelligence data, controlling air traffic, and detecting and identifying targets.

Published tests selected as markers for the induction construct included EAS-6 Numerical Reasoning and ETS Figure Classification. In the Numerical Reasoning Test, subjects are asked to examine a series of numbers to determine the pattern or the principle governing the relationship among the numbers in the series; subjects must then apply the principle to identify the number appearing next in the series. In the ETS Figure Classification Test, subjects are asked to examine two (or three) groups of figures to determine how the figures in one group are alike and how the groups differ; subjects must then classify test figures into one of the two (or three) groups.

Our plan for developing Reasoning Test 1 was to construct a test that was similar to the task appearing in EAS-6 Numerical Reasoning, but with one major difference: items would be composed of illustrations rather than numbers. Test items were constructed to represent varying degrees of difficulty ranging from very easy to very difficult. Following item development, time limits were established to allow sufficient time for subjects to complete all or nearly all items. Thus, Reasoning 1 was designed as a power measure of induction.

Test Development. Reasoning Test 1 items present subjects with a series of four figures. The task is to identify the pattern or relationship among the figures and then to identify from among five possible answers the one figure that appears next in the series. In the original test, subjects were asked to complete 30 items in 14 minutes. Sample items are provided in Figure 3.17.

127

Results from the first tryout, conducted at Fort Carson, indicate that subjects, on the average, completed 29.5 (SD = 1.39) items and obtained a mean score of 20.8 (SD = 3.54). Inspection of difficulty levels indicated that items were unevenly distributed between the two test parts. Items were therefore reordered to ensure that easy and difficult items were equally distributed throughout both test parts. Only minor modifications were made to test items; for example, one particularly difficult item was redrawn to reduce the difficulty level.

Data collected at Fort Campbell indicate that again nearly all subjects completed the test (mean = 29.7, SD = 1.50). Further, test administrators reported that those who completed the test finished early. Thus, the 14-minute time limit was reduced to 12 minutes. Further, two items were revised because distractors yielded higher item-total correlations than the correct response.

Pilot Test Results. Data collected at the third tryout, conducted at Fort Lewis, are reported in Table 3.9. Subjects, on the average, completed 29.4 items with about 84 percent of the subjects completing the entire test. Test scores, computed as the total number correct, ranged from 4 to 29 with a mean of 19.6.

Item difficulty levels ranged from .26 to .92 with a mean of .66. Item-total correlations averaged .45 (SD = .10) with a range of .24 to .60. Part 1 and Part 2 correlate .64. The split-half reliability estimate corrected for test length is .78, while the Hoyt value is .86. These results indicated the test was in good shape; it was a reliable power test with acceptable distributional properties.

One of the marker tests, ETS Figure Classification, was administered at the first two tryout sites. The Fort Carson data indicate Reasoning Test 1 correlates .34 (N = 22) with this measure, while the Fort Campbell data indicate that the two correlate .25 (N = 56). Because the task involved in Reasoning 1 differs from that in ETS Figure Classification, the low value of these correlations is not alarming.

Two other marker measures of induction, SRA Word Grouping and DAT Abstract Reasoning, were administered at the Fort Lewis tryout. These data indicate that Reasoning 1 correlates .47 with Word Grouping and .74 with Abstract Reasoning. These data are compatible with our understanding of these two marker measures of induction. Word Grouping contains a verbal component while Abstract Reasoning measures induction via figural reasoning, similar to Reasoning Test 1.

Modifications for the Fort Knox Field Test. For the Fort Knox field test, instructions for Reasoning Test 1 were revised slightly.

Reasoning Test 2

Development Strategy. This measure was also designed to assess induction using items that require figural reasoning.

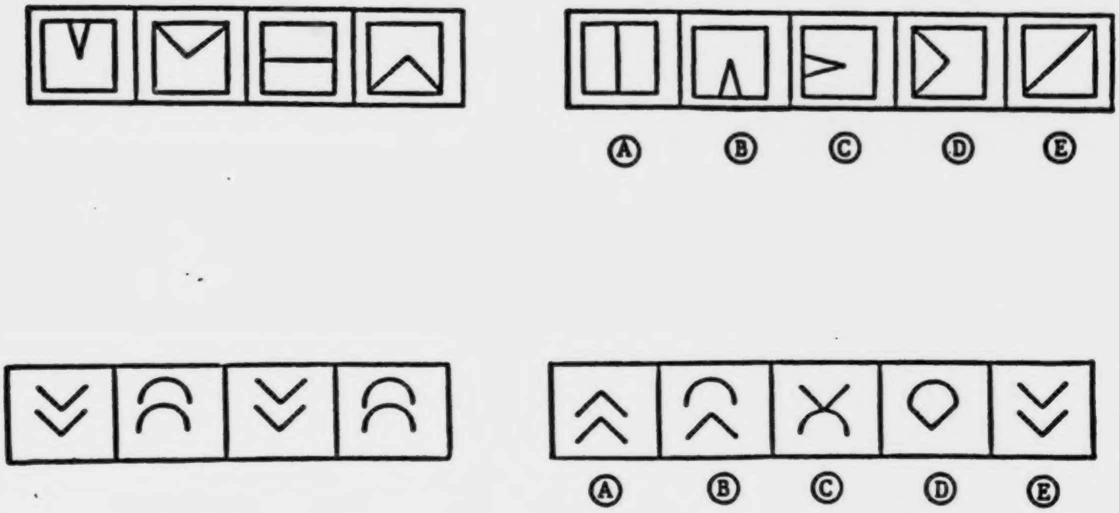

A
B
C
D
E
A
B
C
D
E

Figure 3.17. Sample items from Reasoning Test 1.

Table 3.9

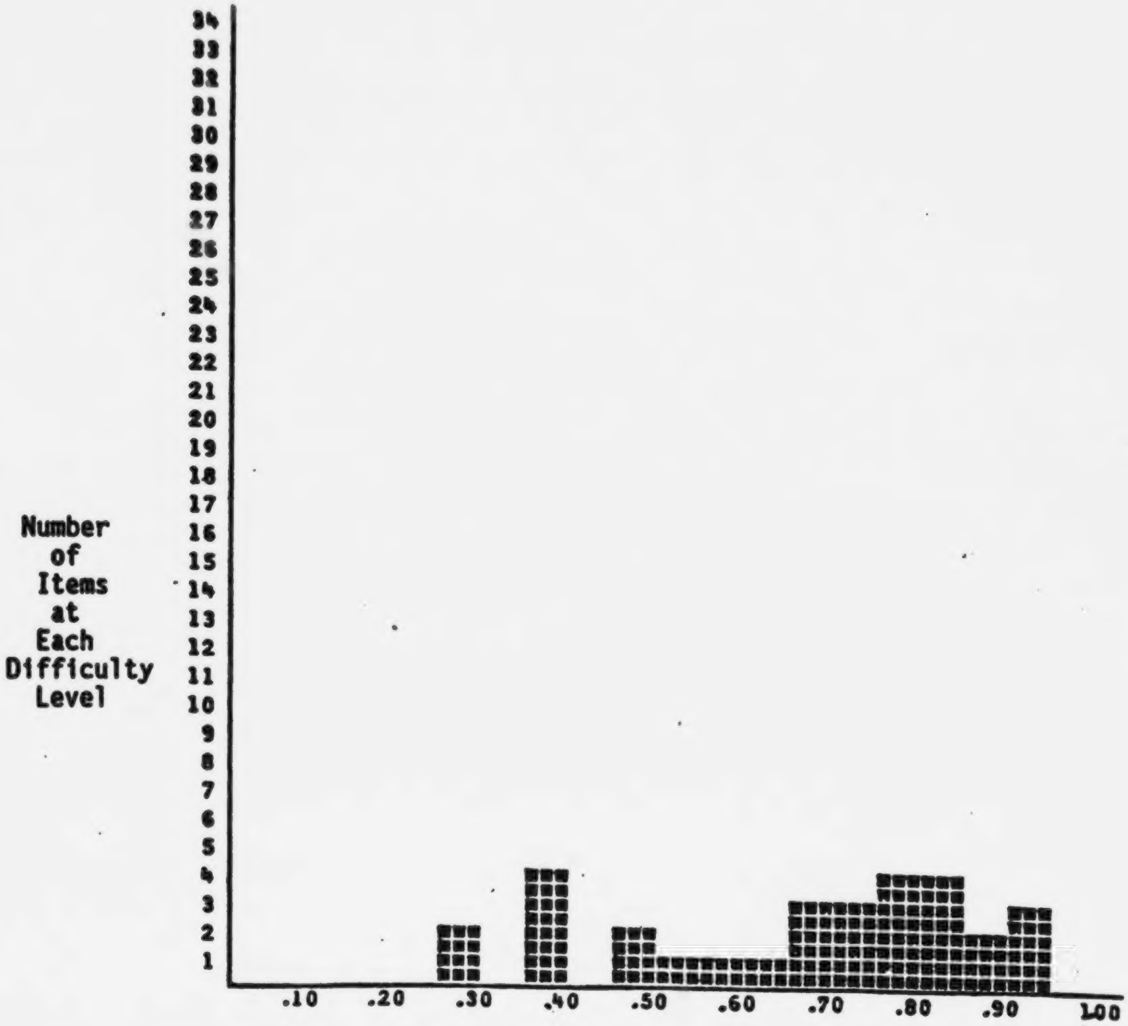
Pilot Test Results from Fort Lewis: Reasoning Test 1

	Total	Part 1	Part 2
Number of Items	30	15	15
Time Allowed (minutes)	12	6	6
Number of Subjects	118	118	118
Number of Items Completed			
Mean	29.44	14.73	14.71
Standard Deviation	1.62	0.87	0.96
Range	22-30	10-15	10-15
Last Item Completed by 80% of the Sample	N/A	15	15
Percentage of Subjects Completing All Items	84%	88%	89%
Number of Items Correct			
Mean	19.64	9.61	10.03
Standard Deviation	5.75	3.16	3.20
Range	4-29	1-15	1-15
Total-Part Intercorrelations			
Total	**	.90	.91
Part 1		**	.64
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .78

Hoyt Internal Consistency = .86

Reasoning Test 1



Item Difficulty Level (Proportion Passing)

Mean = .66

SD = .20

Range = .26 - .92

NOTE: Number of items in the test = 30.

Figure 3.18. Distribution of item difficulty levels: Reasoning Test 1.

Published tests serving as markers for Reasoning Test 2 include EAS-6 Numerical Reasoning and ETS Figure Classification; these measures were described for Reasoning Test 1. The original strategy was to develop Reasoning Test 2 fairly similarly to ETS Figure Classification. Initial Preliminary Battery analyses conducted on ETS Figure Classification data (N = 1,863) indicated that this test was too highly speeded for the target population (Hough, et al., 1984). For example, 80 percent of recruits completing the Figure Classification test finished fewer than half of the 112 items. Further, although item difficulty levels varied greatly, the mean value indicated most items are moderately easy (mean = .73, SD = .22, range = .06 to .98). Thus, although the ETS Figure Classification test served as the marker in early test development planning for Reasoning 2, the new measure differed in several ways, as described below.

First, ETS Figure Classification requires subjects to perform two tasks: to identify similarities and differences among groups of figures, and then to classify test figures into those groups. Items in Reasoning Test 2 were designed to involve only the first task, identifying similarities and differences among figures. Second, test items on Reasoning 2 were constructed to reflect a wide range of difficulty levels, with the average item falling in the moderately difficult range. Finally, because the items would be more difficult overall, we decided that Reasoning 2 would contain fewer items than were included in the Figure Classification Test. The time limit for Reasoning 2 was established to ensure that most subjects would complete the test. Thus, Reasoning 2 was designed as a power measure of figural reasoning, with a broad range of item difficulties.

Test Development. Reasoning 2 test items present subjects with five figures. Subjects are asked to determine which of the four figures are similar in some way, thereby identifying the one figure that differs from the others. (See Figure 3.19.) This test, when first administered, contained 32 items with an 11-minute time limit.

Results from the Fort Carson tryout indicated that nearly all subjects completed the entire test (mean = 31.6, SD = 1.09, N = 38). Item difficulty levels were somewhat higher than expected, ranging from .05 to 1.00 with a mean of .71 (SD = .29). Because eight items yielded item difficulty levels of .97 or above, these items were either modified or replaced to increase item difficulties. Moreover, inspection of item difficulties indicated that Part 1 contained a greater proportion of the easier items, so items were redistributed throughout the test to obtain an equal mix of easy and difficult items, and to attempt to increase the relatively low, part-part correlation (*r* = .32).

For the Fort Campbell tryout, Reasoning 2 again contained 32 items with an 11-minute time limit. Data from this tryout indicated that, for the most part, the test possessed desirable psychometric qualities. For example, nearly all subjects completed the test (mean = 31.1, SD = 1.91). Test scores ranged from 9 to 26 with a mean of 19.1 (SD = 3.56) and the test was a bit more difficult (mean = .56., SD = .34). Although the part-part correlation increased from the first tryout, it still remained low (i.e, Fort Campbell *r* = .40 versus Fort Carson *r* = .32).

A few changes were made in the test prior to the third tryout. For example, four items contained a distractor that was selected more often and

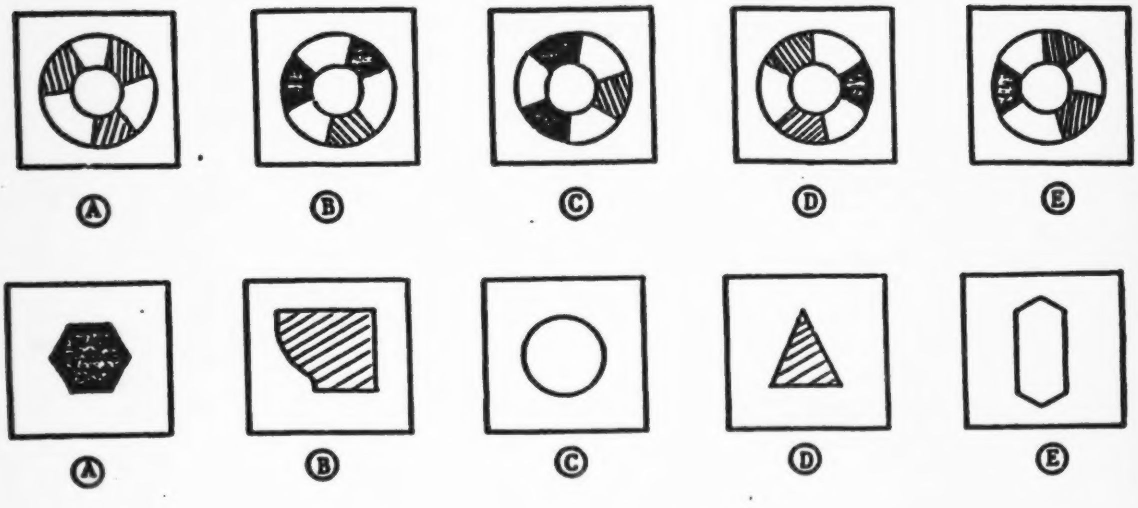


Figure 3.19. Sample items from Reasoning Test 2.

which yielded a higher item-total correlation than the correct response; these distractors were revised. Further, test administrators at Fort Campbell noted that the time limit could be reduced without altering test completion rates. Consequently, the time limit was reduced to 10 minutes.

Pilot Test Results. Results from the third tryout are reported in Table 3.10. Seventy percent completed the entire test, but 84 percent completed the separately-timed first half and 79 percent completed the second half. Thus, these results indicate that the test is probably still a power test (recall our practical rule of thumb was 80 percent completing all items) even with the reduced time limit. Test scores range from 11 to 28 with a mean of 21.8 (SD = 3.38).

Item difficulties range from .17 to 1.00 with a mean of .64 and standard deviation of .19. Item-total correlations averaged .26 (SD = .14) with a range of -.04 to .53. Parts 1 and 2 correlate .46. The split-half reliability estimate, corrected for test length, is .63 while the Hoyt value is .61. These values suggest that this is a more heterogeneous test of figural reasoning than is Reasoning Test 1. These data indicate that the test is acceptable in terms of score distribution, reliability, and power vs. speed continuum.

The marker test, ETS Figure Classification, was administered at the first two tryouts. Correlations between Reasoning 2 and its marker are .35 (N = 30 at Fort Carson) and .23 (N = 56 at Fort Campbell). These low correlations are not too surprising, given the task requirement differences and power versus speed component differences between these two measures. Two other marker measures of induction, SRA Word Grouping and DAT Abstract Reasoning, were administered at the third tryout. These data indicate that Reasoning 2 correlates .48 with Word Grouping and .66 with Abstract Reasoning (N = 118). Once again, these differences in correlations are expected; as noted earlier, Word Grouping contains a verbal component whereas Abstract Reasoning, like Reasoning 2, assesses induction using figural items.

Table 3.10

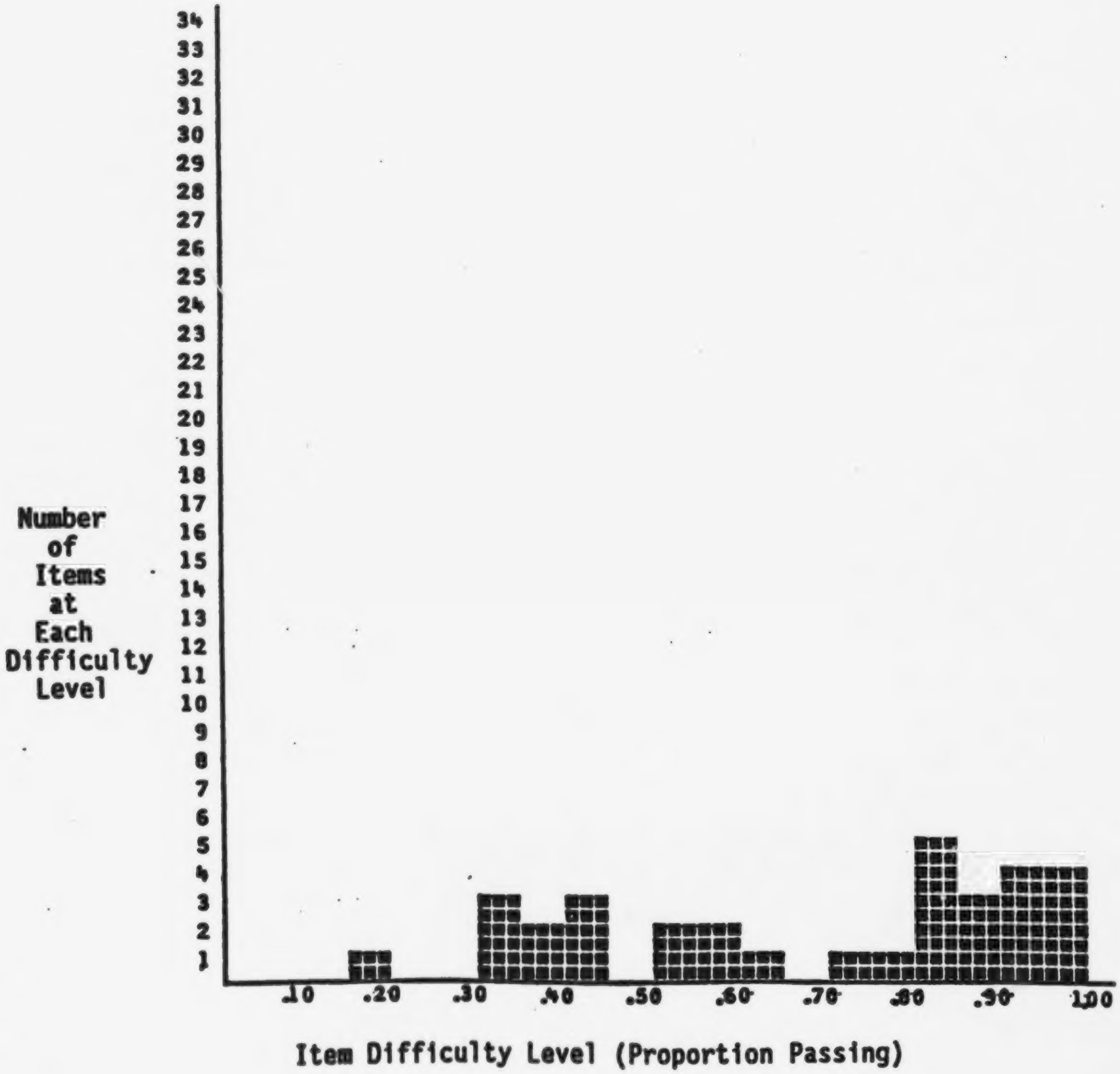
Pilot Test Results from Fort Lewis: Reasoning Test 2

	Total	Part 1	Part 2
Number of Items	32	16	16
Time Allowed (minutes)	10	5	5
Number of Subjects	118	118	118
Number of Items Completed			
Mean	31.19	15.75	15.45
Standard Deviation	1.78	.69	1.38
Range	22-32	12-16	8-16
Last Item Completed by 80% of the Sample	N/A	16	15
Percentage of Subjects Completing All Items	70%	84%	79%
Number of Items Correct			
Mean	21.82	11.31	10.51
Standard Deviation	3.38	1.73	2.21
Range	11-28	7-15	4-15
Total-Part Intercorrelations			
Total	**	.82	.88
Part 1		**	.46
Part 2			**

Split-Half Reliability (Spearman-Brown Corrected) = .63

Hoyt Internal Consistency = .61

Reasoning Test 2



Mean = .64

SD = .19

Range = .17 - 1.00

NOTE: Number of items in the test = 32.

Figure 3.20. Distribution of item difficulty levels: Reasoning Test 2.

__Modifications for the Fort Knox Field Test__. The only change made was in the response format. Reasoning Test 2 contained 32 items with a 10-minute time limit for the Fort Knox field test.

OVERALL ANALYSIS OF PILOT TEST RESULTS FOR COGNITIVE PAPER-AND-PENCIL MEASURES

In this section, we analyze the data available as of August 1984 for the ten cognitive paper-and-pencil measures. This includes a summary of pilot test score information, intercorrelations among the ten measures, results from factor analyses, and data comparing subgroup test scores.

Before providing a summary of the cognitive test data, a word about the source of these data and how they will be used is warranted. As noted, the bulk of the data reported here was obtained from the final pilot test at Fort Lewis tryout. The sample size at Fort Lewis was sufficient for many of the analyses performed (e.g., psychometric characteristics of test response).

For some analyses, however, these data serve as a first step in structuring our understanding of these measures. For example, we provide results from a factor analysis of the intercorrelations among the ten measures. These data provide preliminary information about the underlying structure of the test score data. Another example of tentative conclusions stems from comparisons of subgroup test scores; for the most part, the sample sizes of the subgroups are fairly small and, therefore, results should not be viewed as conclusive.

Table 3.11 summarizes the Fort Lewis data discussed earlier in this chapter. For each measure we include the number of test items, mean test score and standard deviation, mean item difficulty level, and split-half reliability corrected for test length. Note that all data are based on a sample size of 118 with the exception of the Path Test data which is based on a sample size of 116.

Test Intercorrelations and Factor Analysis Results

Table 3.12 contains the intercorrelation matrix for the ten cognitive ability measures. One of the most obvious features of this matrix is the high level of correlations across all measures. The correlations across all test pairs range from .40 to .68. These data suggest that the test measures overlap in the abilities assessed.

This finding is not altogether surprising. For example, four of the ten measures were designed to measure spatial abilities such as visualization, rotation, and scanning. The Shapes Test, designed to measure field independence, also includes visualization components. The three tests constructed to measure spatial orientation involve visualization and rotation tasks. The final two measures, Reasoning Test 1 and Reasoning Test 2, also require visualization at some level to identify the principle governing relationships among figures and to determine the similarities and differences among figures. Thus, across all measures, abilities needed to complete the required tasks overlap to some degree. This overlap is demonstrated in the intercorrelation matrix.

To enable a better understanding of the similarities and differences among these measures or the underlying structure of these measures, the

Table 3.11

Cognitive Paper-and-Pencil Measures:
Summary of Fort Lewis Pilot Test Results

Measure	No. of Items	Mean Score	SD	Mean Item-Difficulty Levels	Split-Half* $r_{xx'}$
SPATIAL VISUALIZATION					
Rotation					
Assembling Objects	40	28.14	7.51	.70	.79
Object Rotation	90	73.36	15.40	.82	.86
Scanning					
Path	44	28.28	9.08	.64	.82
Mazes	24	19.30	4.35	.80	.78
FIELD INDEPENDENCE					
Shapes	54	29.28	9.14	.54	.82
SPATIAL ORIENTATION					
Orientation 1	150	117.86	24.16	.79	.92
Orientation 2	24	11.53	6.20	.48	.89
Orientation 3	20	8.71	5.78	.44	.88
REASONING					
Reasoning 1	30	19.64	5.75	.66	.78
Reasoning 2	32	21.82	3.38	.64	.63

*All reliability estimates (split-halves with part 1-part 2 separately timed) have been corrected with the Spearman-Brown procedures.

Table 3.12

Intercorrelations Among the Ten Cognitive Paper-and-Pencil Measures[a]

Measure	1. Assembling Objects	2. Object Rotation	3. Path	4. Maze	5. Shapes	6. Orientation 1	7. Orientation 2	8. Orientation 3	9. Reasoning 1	10. Reasoning 2
1. Assembling Objects	--									
2. Object Rotation	.53	--								
3. Path	.52	.45	--							
4. Maze	.59	.57	.60	--						
5. Shapes	.61	.50	.51	.56	--					
6. Orientation 1	.62	.52	.54	.52	.56	--				
7. Orientation 2	.60	.45	.48	.51	.47	.53	--			
8. Orientation 3	.62	.50	.40	.47	.60	.68	.65	--		
9. Reasoning 1	.62	.52	.60	.58	.59	.59	.56	.60	--	
10. Reasoning 2	.53	.50	.48	.52	.51	.54	.53	.53	.63	--

---

[a] All correlations are computed from a sample size of 118 except those involving the Path Test which are based on sample size of 116.

intercorrelation matrix was factor analyzed. A principal factors extraction was performed with iterated, squared multiple correlations as the communality estimates. Several solutions were computed, ranging from two to five factors. The rotated orthogonal solution for four factors appeared most psychologically meaningful. Results from this solution appear in Table 3.13.

As shown in the table, to interpret results from the four-factor solution we first identified all factor loadings of .35 or higher. Next, we examined the factor loading pattern for each measure and then identified measures with similar patterns to form test clusters. Five test clusters or groups, labeled A through E, are identified in Table 3.13. These clusters represent a first attempt to identify the underlying structure of the cognitive measures included in the Pilot Trial Battery. Each test cluster is described below:

Group A - Assembling Objects and Shapes Tests. Recall that the Shapes Test requires the subject to locate or disembed simple forms from more complex patterns, while the Assembling Objects Test requires the subject to visualize how an object will appear when its components are put together. Both measures require subjects to visualize objects or forms in new or different configurations. Further, these measures contain both power and speed components with each falling more toward the speed end of the continuum.

Group B - Object Rotation, Path, and Maze Tests. Object Rotation involves two-dimensional rotation of objects or forms while the Path and Maze tests involve visually scanning a map or diagram to identify the best pathway or the one pathway that leads to an exit. These measures are all highly speeded; that is, subjects are required to perform the tasks at a fairly rapid rate. Further, the tasks involved in each of these measures appear less complex or easier than those involved in the Assembling Objects or Shapes tests.

Group C - Orientation 1 and Orientation 3 Tests. Orientation Test 1 requires one to compare compass directions provided on a test circle and a Given Circle, while Orientation Test 3 involves using a map, compass directions, and present location to determine which direction to go to reach a landmark on the map. Both measures require a subject to quickly and accurately orient oneself with respect to directions on a compass and landmarks in the environment despite shifts or changes in the directions. Both are highly speeded measures of spatial orientation.

Group D - Orientation Test 2. This measure involves mentally rotating a frame so that it corresponds to or matches up with the picture inside, and then visualizing how components on the frame (a circle with a dot) will appear after it has been rotated. This appears to be a very complex spatial measure that requires several abilities such as visualization, rotation, and orientation. In addition to the task complexity differences, this measure may also differ from other spatial measures on the power-speed continuum. Unlike the other spatial measures included in the Pilot Trial Battery, Orientation 2 is a power rather than a speed test.

Group E - Reasoning 1 and Reasoning 2 Tests. Reasoning Test 1 re-

Table 3.13

Rotated Orthogonal Factor Solution for Four Factors[a]

	I		II		III		IV		h^{2b}
Shapes	.47		.49←	A					.568
Assembling Objects	.47		.48←						.621
Object Rotation	.50←	B	.37						.473
Path	.55←				.40				.541
Maze	.76←								.727
Orientation 1	.39		.57←	C					.617
Orientation 3			.79←				.35		.827
Orientation 2			.35				.74 ←	D	.684
Reasoning 1	.39		.35		.67←	E			.778
Reasoning 2	.37		.36		.44←				.521

[a] Factor loadings of .35 or higher are shown.

[b] h^2 = Proportion of total test score variance in common with other tests, or common variance.

quires one to identify the principle governing the relationship or pattern among several figures, while Reasoning Test 2 involves identifying similarities among several figures to isolate the one figure that differs from the others. As noted above, these measures appear to involve visualization abilities. The reasoning task involved in each, however, distinguishes these measures from the other tests included in the Pilot Trial Battery.

Results from analyses of the Fort Lewis data provide a preliminary structure for the cognitive paper-and-pencil tests designed for the Pilot Trial Battery. Correlations among the measures indicate that all measures require spatial visualization abilities at some level. The measures may, however, be distinguished by the type of task, task complexity, and speed and power component differences.

Subgroup Analyses Results

Mean test scores were compared for two pairs of subgroups: (a) blacks and whites, and (b) males and females. The sample sizes for each subgroup are fairly small with the exception of the male subsample (N = 97). Consequently, reported differences are intended to provide only a "ball-park" estimate of the mean effect size differences between the subgroups. It is important to note that the reported subgroup differences may, in fact, be inaccurate estimates of the true differences in the target population. This may occur for several reasons, such as restriction in range of test score data due to selection, and primarily, sampling error because of the small samples used here.

Table 3.14 contains the mean effect size differences for blacks and whites on the various tests. The differences for these groups range from .63 to 1.17. Note that the largest differences appear in Orientation Test 1 (mean effect size = 1.17), Assembling Objects Test (mean effect size = 1.10), and the Shapes Test (mean effect size = 1.06). The smallest differences appear for Object Rotation Test (mean effect size = .63) and Reasoning Test 2 (mean effect size = .72). These differences are in line with the size of white-black differences usually found with cognitive, paper-and-pencil tests.

Table 3.15 contains mean effect size differences for males versus females on each of the ten measures. Mean effect size differences range from .05 to .87. The largest difference appears for the Object Rotation Test while the smallest difference appears for Orientation Test 2. These gender differences represent values somewhat lower than those usually found in the literature, indicating that they may be underestimates for the target population.

Once again, however, we emphasize strongly that these results are suggestive only, due to the small sample sizes.

Other Cognitive Tests

In this chapter we have focused on the cognitive paper-and-pencil measures. Other cognitive measures were administered in the Pilot Trial Battery; those measures were administered via computer and are described in Chapter 5. Correlations among the cognitive paper-and-pencil tests and the cognitive computer tests are also reported in that chapter. Before de-

Table 3.14

Subgroup Analyses of Cognitive Paper-and-Pencil Tests: White-Black Mean Score Differences in Pilot Test

		Whites			Blacks			
Construct & Test	No. Possible	N	Mean	SD	N	Mean	SD	Mean[a] Effect Size
SPATIAL VISUALIZATION (Rotation)								
Assembling Objects	40	66	30.85	5.80	30	23.47	8.37	1.10
Object Rotation	90	66	77.00	12.54	30	67.97	17.65	.63
SPATIAL VISUALIZATION (Scanning)								
Path	44	65	30.35	8.80	30	22.97	8.84	.84
Maze	24	66	20.58	3.88	30	16.57	4.31	1.00
FIELD INDEPENDENCE								
Shapes	54	66	33.03	8.31	30	24.50	7.37	1.06
SPATIAL ORIENTATION								
Orientation 1	150	66	127.65	19.54	30	104.00	21.89	1.17
Orientation 2	24	66	13.33	6.35	30	8.53	4.98	.81
Orientation 3	20	66	10.80	5.43	30	6.20	5.13	.86
REASONING								
Reasoning 1	30	66	21.53	5.12	30	17.17	5.56	.83
Reasoniing 2	32	66	22.73	3.46	30	20.23	3.56	.72

[a] Mean effect size = $$\frac{\text{Mean (Whites) - Mean (Blacks)}}{\text{Pooled Standard Deviation}}$$

This statistic provides an estimate of the difference in test score performance expressed in standard deviation units.

Table 3.15

Subgroup Analyses of Cognitive Paper-and-Pencil Tests: Male-Female Mean Score Differences in Pilot Test

---

Construct & Test	No. Possible	Males N	Males Mean	Males SD	Females N	Females Mean	Females SD	Mean[a] Effect Size
SPATIAL VISUALIZATION (Rotation)								
Assembling Objects	40	97	28.43	7.68	21	26.81	6.47	.22
Object Rotation	90	97	75.63	14.37	21	62.90	15.67	.87
SPATIAL VISUALIZATION (Scanning)								
Path	44	95	28.62	9.55	21	26.76	6.29	.21
Maze	24	97	19.80	4.13	21	16.95	4.57	.68
FIELD INDEPENDENCE								
Shapes	54	97	29.82	9.07	21	26.76	8.99	.34
SPATIAL ORIENTATION								
Orientation 1	150	97	119.01	24.47	21	112.52	21.93	.27
Orientation 2	24	97	11.59	6.28	21	11.29	5.85	.05
Orientation 3	20	97	8.93	5.65	21	7.71	6.27	.21
REASONING								
Reasoning 1	30	97	19.76	5.63	21	19.05	6.26	.12
Reasoniing 2	32	97	21.91	3.76	21	21.43	2.32	.14

---

[a] Mean effect size = $$\frac{\text{Mean (Males) - Mean (Females)}}{\text{Pooled Standard Deviation}}$$

This statistic provides an estimate of the difference in test score performance expressed in standard deviation units.

scribing the computer-administered tests, we provide results from the field test analyses of the paper-and-pencil cognitive measures in Chapter 4.

Chapter 3 References

Guilford, J. P., & Lacey, J. I. (Eds.) (1947). Printed classification tests. *Army Air Forces Aviation Psychology Research Program Reports*, *5*, Washington, DC: Government Printing Office.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J. S., & Peterson, N. G. (1984). *Covariance analyses of cognitive and non-cognitive measures of Army recruits: An initial sample of Preliminary Battery Data*. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada. In Eaton, N. K. et al. (Eds.) (1984), ARI Technical Report 660.

Wing, H., Peterson, N. G., & Hoffman, E. G. (1984). *Expert judgments of predictor-criterion validity relationships*. Paper presented at the 92nd Annual Convention of the American Psychological Assocation, Toronto, Canada. In Eaton, N. K. et al. (Eds.) (1984), ARI Technical Report 660.

CHAPTER 4

COGNITIVE PAPER-AND-PENCIL MEASURES: FIELD TEST

Marvin D. Dunnette, VyVy A. Corpe, and Jody L. Toquam

In this chapter we describe analyses of the field test of the cognitive paper-and-pencil tests in the Pilot Trial Battery, administered at Fort Knox in September 1984. The procedures and sample for this field test were described in Chapter 2. In this chapter we present descriptive statistics for the tests, internal consistency and test-retest reliabilities, an analysis of gains in scores when the tests are taken a second time, and analyses of the relationships between the ASVAB subtests and the Pilot Trial Battery cognitive tests. Later chapters of this report will extend analysis of the data from the field tests to cover the relationships of the cognitive paper-and-pencil measures with the other measures--computer-administered perceptual/psychomotor, and non-cognitive paper-and-pencil--which were also part of the Pilot Trial Battery. We note here that parts of this chapter are drawn from Toquam et al. (1985).

A concise description of each of the ten tests, along with a sample item or items from each test, is contained in Figure 4.1. Copies of the full Pilot Trial Battery as administered at Fort Knox are contained in Appendix G.

ANALYSES OF DATA FROM FIELD TEST ADMINISTRATION

Mean Scores and Reliability Estimates

Table 4.1 shows the means, standard deviations, and three estimates of the reliabilities of the cognitive tests administered in the field test of the Pilot Trial Battery. The means and standard deviations are similar to the results obtained at the last pilot test at Fort Lewis (see Table 3.11), except for two tests. The mean score for Object Rotation is about 14 points lower for the field test (59.62 vs. 73.36), but this was expected and intended since we had decreased the time allowed on this test from 8 minutes to 7.5 minutes--in order to avoid a possible ceiling effect. Orientation Test 1 also showed a mean score decrease, from 117.86 to 88.65. No changes had been made in the test so it is not clear why this occurred. The decrease is not alarming, however, since the examinees still answered about .59 of the items correctly which is in the range of test difficulty we desired (about .50 to .70) for this set of tests.

Difficulty levels for the other tests are also in this .50 to .70 range, except for Orientation 3. (The test difficulties are not shown in Table 4.1 but can be obtained by dividing the mean score by the total number of items.) This test appears to be a bit more difficult than desired (difficulty = .39), but this appears not to adversely affect the test score variance (standard deviation = 5.68) or its reliability (split half reliability = .88 and test-retest reliability = .84).

Three estimates of reliability are shown in Table 4.1. The first one, labeled split-half, is actually computed on the Fort Lewis pilot test data, not on the Fort Knox field test data. Separately timed halves were administered at Fort Lewis, but time limitations did not allow this at Fort Knox. We have included these estimates because they are more appropriate than coefficient Alpha for those tests that are moderately or highly speeded. All of the PTB cognitive tests are at least moderately speeded, except Orientation 2, Reasoning 1, and Reasoning 2.

Examination of these reliability estimates shows that all of the tests are acceptably reliable, with the possible exception of Reasoning 2. The estimates of internal consistency (split half and coefficient Alpha) are .78 or higher, except for Reasoning 2 and the test-retest reliability estimates (two-week interval) are .64 or higher, except for this test.

Gain Score Analysis

The collection of retest data allowed us the opportunity to examine the extent to which test score distributions might change when the tests are taken a second time. Generally speaking, prior exposure to a test leads to an increase in test scores, especially if the exposure is very close to the time the test is taken. In this case, the soldiers completed all the cognitive tests twice, with a two-week interval between administrations.

Our concern was that taking the test a second time might lead to a large increase in scores. If so, this would need to be taken into account if the tests were used in an operational setting. (Retest opportunities could be controlled or limited, or parallel forms could be developed.)

149

Cognitive Paper-and-Pencil Measures

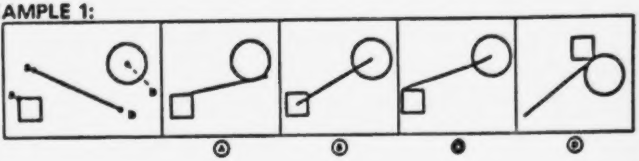
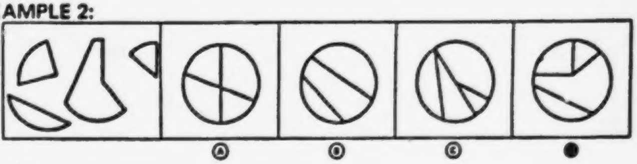
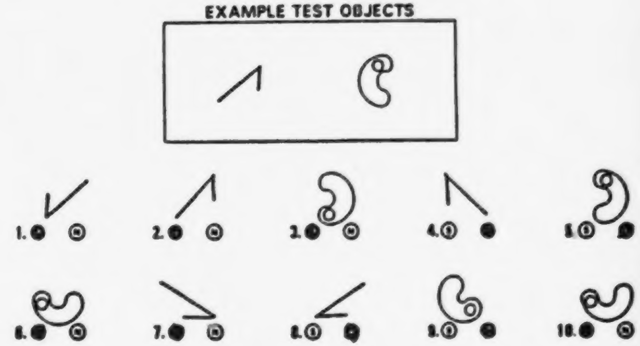
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
SPATIAL VISUALIZATION - ROTATION		
Assembling Objects	The test contains 40 items with a 16 minute time limit. The subject's task involves figuring out how an object will look when its parts are put back together again. There are two types of problems in the test. In one part, the item shows a picture of labelled parts. By matching the letters, it can be "seen" where the parts should touch when the object is put together correctly. The second type of problem does not label any of the parts. The parts fit together like the pieces of a puzzle. In each section, four possible figures are provided and the subject must pick the correct one.	EXAMPLE 1:  EXAMPLE 2: 
Object Rotation	The test contains 90 items with a 7 1/2 minute time limit. The subject's task involves examining a test object and determining whether the figure represented in each item is the same as the object, only rotated, or is not the same as the test object (e.g., flipped over). For each test object there are 5 test items, each requiring a response of "same" or "not same".	EXAMPLE TEST OBJECTS 

Figure 4.1. Description of Cognitive Paper-and-Pencil Measures in Field Test (Page 1 of 4)

Cognitive Paper-and-Pencil Measures

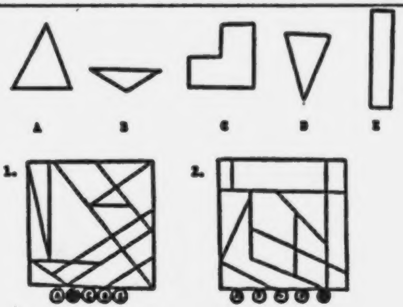
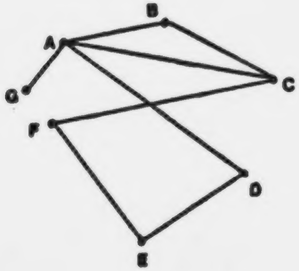
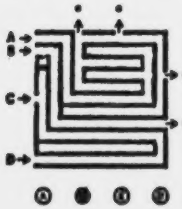
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
SPATIAL VISUALIZATION - FIELD INDEPENDENCE		
Shapes	The test contains 54 items with a 16-minute time limit. At the top of each test page are five simple shapes; below these shapes are six complex figures. Subjects are instructed to examine the simple shapes and then to find the one simple shape located in each complex figure.	
SPATIAL VISUALIZATION - SCANNING		
Path	The test contains 44 items with an 8-minute time limit. Subjects are required to determine the best path or route between two points. Subjects are presented with a map of airline routes or flight paths. The subject's task is to find the "best" path or the path between two points that requires the fewest number of stops.	 The route from: Number of Stops: 1. A to F ● ② ③ ④ ⑤ 2. G to E ① ② ③ ④ ⑤ 3. C to D. ① ② ③ ④ ⑤ 4. G to F ① ② ③ ④ ⑤
Mazes	The test contains 24 items with a 5 1/2 minute time limit. Each item is a rectangular maze with four labelled entrance points and four exit points. The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points.	

Figure 4.1. Description of Cognitive Paper-and-Pencil Measures in Field Test (Page 2 of 4)

Cognitive Paper-and-Pencil Measures

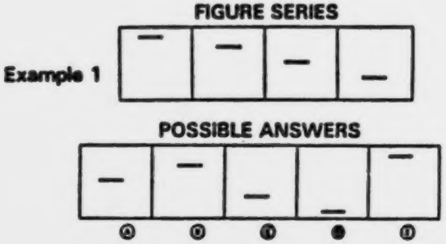
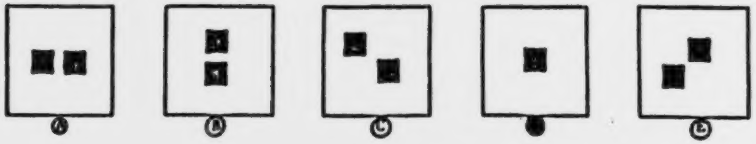
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
INDUCTION		
Reasoning 1	The test contains 30 items with a 12 minute time limit. Subjects are presented with a series of four figures. The task is to identify the pattern or relationship among the figures and then to identify from among five possible answers the one figure that appears next in the series.	
Reasoning 2	The test contains 32 items with a 10 minute time limit. Subjects are presented with five figures. They are then asked to determine which of the four figures are similar in some way, thereby identifying the one figure that differs from the others.	Example 1: 

Figure 4.1. Description of Cognitive Paper-and-Pencil Measures in Field Test (Page 3 of 4)

Cognitive Paper-and-Pencil Measures

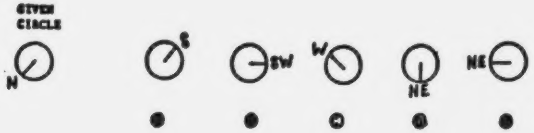
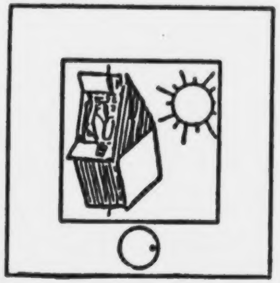
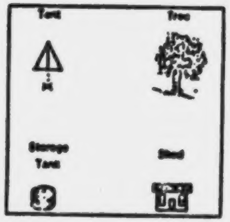
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
SPATIAL ORIENTATION		
Orientation 1	The test contains 150 items (30 5-item sets) with a 10 minute time limit. Each set presents subjects with 6 circles. The first, the Given Circle, indicates the compass direction for North. For most items, North is rotated out of its conventional position (e.g., the top of the circle does not necessarily represent North). Compass directions also appear on the remaining five circles. The subject's task is to determine for each circle, whether or not the direction indicated is correctly positioned by comparing it to the direction of North in the Given Circle.	
Orientation 2	The test contains 24 items with an 10 minute time limit. Each item contains a picture within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture or scene is not in an upright position. The task is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture. After doing so, the subject must then decide where the dot will appear in the circle.	
Orientation 3	The test contains 20 items with a 12 minute time limit. Subjects are presented with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake, and so on. Within each item, subjects are provided with compass directions by indicating the direction of one landmark to another, such as "the forest is North of the camp-site". Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark.	 1. The shed is due north of the tree. You are at the storage tank. Which direction must you travel to reach the tent?

Figure 4.1. Description of Cognitive Paper-and-Pencil Measures in Field Test (Page 4 of 4)

Table 4.1

Means, Standard Deviations, and Reliability Estimates for the Fort Knox Field Test of the Ten Cognitive Paper-and-Pencil Tests

Test	No. Items	Time Allotted (in minutes)	Score Mean[a]	SD[a]	Reliability Coefficients[b]		
					Split Half (N = 118)	Coefficient Alpha (N = 290)	Test-Retest (N = 97 to 126)
Assembling Objects	40	16	26.45	8.67	.79	.92	.74
Object Rotation	90	7.5	59.62	18.98	.86	.97	.75
Path	44	8	26.37	8.86	.82	.92	.64
Maze	24	5.5	17.76	4.45	.78	.89	.71
Shapes	54	16	26.39	10.21	.82	.92	.70
Orientation 1	150	10	88.65	34.74	.92	.98	.67
Orientation 2	24	10	11.46	5.96	.89	.88	.80
Orientation 3	20	12	7.73	5.68	.88	.90	.84
Reasoning 1	30	12	19.57	5.23	.78	.83	.64
Reasoning 2	32	10	21.60	3.63	.63	.65	.57

[a] Ns range from 292 to 298 for mean and SD calculations.

[b] The split-half coefficient is computed on pilot test data from Fort Lewis, where two separately timed halves were given, and is corrected to full test length. Coefficient alphas are based on the Fort Knox data and are overestimates for the speeded tests. The test-retest interval was two weeks.

Table 4.2 shows the gain scores for persons in the retest sample. Four of these tests showed gain scores that appeared to be higher than we thought desirable: Shapes, Orientation 1, Path, and Object Rotation. In order to estimate the seriousness of this concern we located gain scores for a number of other cognitive tests that measured similar constructs. We found that gain scores of similar magnitude occurred on those tests as well (e.g., on General Aptitude Test Battery tests of spatial aptitude and form perception, gain scores ranged from .46 to .62, U.S. Department of Labor, 1970)). Although this finding did not solve the concern with these relatively large, undesirable gain scores, it did indicate that gain scores of this magnitude are not uncommon for tests of this type.

Inspection of the last two columns in Table 4.2 indicated that much of the gain probably occurred because the soldiers attempted more items the second time they took the test. This is certainly to be expected since the retested soldiers would be more familiar with item types and instructions.

The gain score analysis showed that persons could, on the average, increase their scores on several of the PTB cognitive tests to a degree that seems to be cause for some concern in an operational setting. However, a brief review of the literature showed that gain scores of the magnitude we found were also found for commonly used, published tests of the same type. This indicates that our evaluation of the need for concern may be unduly high.

Covariance with ASVAB Subtests

One of the primary goals, and criteria for evaluation of our success, was the development of new predictor measures that would complement the ASVAB rather than measure the same things (see Chapter 1 for a discussion of the overall strategy of predictor development). In order to evaluate our progress toward that goal, we analyzed the covariance of the Pilot Trial Battery with the ASVAB. In this section we report the correlations between these measures and a statistic, called uniqueness, that indicates the amount of overlap between one test and a set of other tests.

We take up the correlations first. If we had achieved our goal of complementing the ASVAB, then the PTB cognitive tests should correlate low to moderately with the ASVAB subtests.

Table 4.3 contains the intercorrelations for the ASVAB subtests and the cognitive paper-and-pencil measures. Note that we have also included scores on the Armed Forces Qualification Test (AFQT). These correlations are based on the Fort Knox field test sample, but include only those subjects with test scores available on all variables (N = 168).

In examining these relationships, we first looked at the correlations between tests within the same battery. Correlations between ASVAB subtest scores range from .02 to .74 (absolute values). The range of intercorrelations is a bit more restricted when examining the relationships between the cognitive paper-and-pencil test scores (.27 to .67). This range of values reflects the fact that the Pilot Trial Battery measures were designed to tap fairly similar cognitive constructs.

Table 4.2

Gains on Pilot Trial Battery Cognitive Tests for Persons Taking Tests at Both Time 1 and Time 2

Test	No. Items	No. Subjects	Time 1 Mean	Time 1 SD	Time 2 Mean	Time 2 SD	Gain[a]	Items Attempted by 75% of Subjects Time 1	Items Attempted by 75% of Subjects Time 2
Assembling Objects	40	113	25.68	9.13	28.23	8.84	0.28	32	40
Object Rotation	90	125	61.23	19.60	71.34	15.92	0.57	55	69
Path	44	126	27.43	8.43	32.46	7.83	0.62	28	36
Maze	24	97	17.47	4.28	18.52	4.34	0.24	17	19
Shapes	54	121	27.30	10.71	34.43	11.50	0.64	30	42
Orientation 1	150	123	91.8	33.05	112.49	32.01	0.63	85	110
Orientation 2	24	116	11.64	5.99	12.31	6.12	0.11	24	24
Orientation 3	20	117	7.71	5.63	8.11	5.60	0.08	16	19
Reasoning 1	30	117	20.35	5.03	21.15	5.49	0.15	30	30
Reasoning 2	32	121	21.22	3.76	21.88	3.49	0.17	32	32

[a] Gain = $$\frac{M_2 - M_1}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}}$$

Table 4.3

Intercorrelations Among the ASVAB Subtests and the Cognitive Paper-and-Pencil Measures in the Pilot Trial Battery: Fort Knox Sample
(N = 168)

		AFQT Score	Gen Scienc	Arith Reas	Word Know	Parg Comp	Numb Ops	Code Spd	Auto/Shop	Math Know	Mech Comp	Elec Info	Assmbl Obj	Obj Rotat	Shapes	Maze	Path	Reas 1	Reas 2	Orient 1	Orient 2	Orient 3
ASVAB	AFQT Score																					
	Gen Scienc	61																				
	Arith Reas	87	54																			
	Word Know	81	66	61																		
	Parg Comp	69	43	53	58																	
	Numb Ops	44	02	21	06	14																
	Code Spd	30	-02	14	10	14	55															
	Auto/Shop	45	54	50	45	29	-08	-07														
	Math Know	76	60	74	62	54	19	15	43													
	Mech Comp	55	50	62	54	36	-10	-06	64	57												
	Elec Info	56	66	56	55	39	-03	01	71	59	63											
PTB -- Paper-and- Pencil Tests	Assmbl Obj	44	38	48	40	29	-01	19	39	48	57	38										
	Obj Rotat	30	20	33	18	18	18	13	27	29	36	32	47									
	Shapes	35	29	33	28	20	14	23	17	36	35	26	51	50								
	Maze	32	23	38	16	12	19	19	34	36	44	35	60	58	47							
	Path	35	20	37	21	17	24	31	33	35	42	32	54	59	50	63						
	Reas 1	46	33	48	41	33	04	15	29	47	51	32	66	38	41	51	52					
	Reas 2	41	32	44	34	34	02	12	18	49	38	30	52	27	44	33	37	50				
	Orient 1	45	41	45	35	29	14	20	40	47	50	39	52	55	54	55	58	49	41			
	Orient 2	50	29	52	38	33	18	13	36	48	49	35	51	38	42	49	42	48	47	55		
	Orient 3	59	51	62	48	43	05	15	55	60	63	54	61	42	42	51	46	49	53	67	56	

Examining the correlations between the ASVAB subtests and the PTB cognitive paper-and-pencil tests, we find that the correlations range from -.01 (Assembling Objects and Number Operations) to .63 (Orientation 3 and Mechanical Comprehension). The mean correlation is .33 (SD = .14). Note that across all PTB paper-and-pencil tests, ASVAB Mechanical Comprehension appears to correlate the highest with the new tests. Across all ASVAB subtests, Orientation 3 yields the highest correlations.

These results show that our goal of complementing the ASVAB has largely been achieved. Certainly, the ASVAB subtests and PTB tests are correlated, but not highly. As noted above, the mean correlation is .33 which is moderate for the average correlation between paper-and-pencil, cognitive tests. This complementary nature of the PTB is shown even more straightforwardly by the uniqueness analyses.

Uniqueness Estimates of Cognitive Tests

Table 4.4 shows uniqueness estimates for the ten cognitive paper-and-pencil tests. Uniqueness is estimated by subtracting the squared multiple regression of a set of tests (in this case the ASVAB or PTB) from the reliability estimate for the test of interest ($u^2 = R_{xx} - R^2$). [See Wise and Mitchell (1985) for discussion of this estimate.] Uniqueness is, then, the amount of reliable variance for a test not shared with the tests against which it has been regressed.

The hope was that the PTB tests would have high uniqueness when regressed against the ASVAB. Such results would indicate that the PTB tests complement the ASVAB when all of the ASVAB subtests are taken into account simultaneously, and that the necessary condition for incrementing the ASVAB validity (against job performance) would be present. As Table 4.4 shows, the uniqueness estimates for the PTB when regressed against the ASVAB subtests ranged from .34 (Orientation 3) to .67 (Object Rotation). These estimates are encouraging since there is ample room for incremental validity to occur.

We point out, however, that the ASVAB tests and PTB tests were not administered concurrently. The ASVAB was taken prior to time of entry into the service and the PTB tests were administered to the soldiers about one-and-one-half years, on the average, after they entered the service. This non-concurrent administration operates to reduce the correlation between the two sets of tests, but to an unknown degree. Thus, these uniqueness estimates are overestimates by some unknown amount.

Table 4.4 also shows the R^2 and U^2 for each PTB test when regressed against all the other PTB tests. These U^2 values were expected to be much lower than the U^2 values obtained by regressing each PTB test against the ASVAB subtests, since the PTB tests measure constructs more similar to each other than the constructs in the ASVAB; indeed, they are about 10 to 20 points lower, except for Orientation 3 which is only 4 points lower.

The results of the analyses of the covariance of ASVAB with PTB show that there is moderate overlap between the two batteries. There appears to be a relatively large amount of reliable variance in the PTB cognitive tests that is not accounted for by the ASVAB. This is the necessary condition that must be obtained in order to increment the validity of ASVAB for

Table 4.4

Uniqueness Estimates for Cognitive Tests in Pilot Trial Battery (PTB) Against Tests in PTB and Against Tests in ASVAB

		Other PTB Tests		ASVAB Tests	
Test	Split Half	R^{2*}	U^{2**}	R^{2*}	U^{2**}
Assembling Objects	.79	.59	.20	.40	.39
Object Rotation	.86	.42	.44	.19	.67
Path	.82	.51	.31	.29	.53
Maze	.78	.46	.32	.25	.53
Shapes	.82	.39	.43	.19	.63
Orientation 1	.92	.58	.34	.36	.56
Orientation 2	.89	.45	.44	.30	.59
Orientation 3	.88	.58	.30	.54	.34
Reasoning 1	.78	.45	.33	.29	.53
Reasoning 2	.63	.37	.26	.26	.37

*The R^2 with the other cognitive paper-and-pencil tests and with the ASVAB subtests are the squared multiple regression coefficients corrected for shrinkage using the standard procedure in the Statistical Analysis System (SAS) software package.

**Uniqueness estimates (U^2) were computed using the split-half reliability estimate. The uniqueness is equal to the split-half reliability minus the R^2 with the ASVAB or with other paper-and-pencil tests.

job performance.

Summary of Analyses

The field test analyses showed that the PTB cognitive tests were, for the most part, in excellent shape. The tests have adequate to excellent score distributions and reliabilities, with one test having marginal reliability (Reasoning 2). Four of the ten tests appeared to be susceptible to large increases in test scores when they are taken a second time, but apparently no more so than commonly used published tests. Finally, the PTB cognitive tests do appear to complement the ASVAB, and possess enough reliable score variance that is uncorrelated with ASVAB to allow the possibility of substantial incremental validity for job performance.

As we noted in the opening of this chapter, the relationships of the PTB cognitive, paper-and-pencil tests to other parts of the Pilot Trial battery are covered in later chapters of this report.

CHAPTER 5

PERCEPTUAL/PSYCHOMOTOR COMPUTER-ADMINISTERED MEASURES: PILOT TESTING

Rodney L. Rosse, Norman G. Peterson, Jeffrey J. McHenry,
Jody L. Toquam, Janis S. Houston, and Teresa L. Russell

GENERAL

In Chapter 1 (see *Computer Battery Development*), we provided a description of the early development of the computer-administered measures. We focused on site visits to military laboratories to investigate other efforts to develop computer-administered tests, choice of appropriate hardware, acquisition of appropriate hardware, choice of appropriate computer languages, and a strategy for melding the efforts of programming the computer with the input of staff scientists responsible for developing the various tests.

In that chapter we briefly described early tryouts of the computer-administered measures at the Minneapolis Military Entrance Processing Station and at the Fort Carson pilot test. We add here that these early tryouts focused primarily on (1) making sure the computer programming was working correctly, (2) the general reactions of soldiers to a computer-administered battery, especially the test instructions, and (3) the general effectiveness of commercially available equipment (keyboards and "computer game" joysticks) for acquiring examinee responses. Actual analysis of the test responses themselves was secondary during that phase of the research, however, we learned much that shaped the way the tests were programmed, the instructions and items that were presented, and the way responses were acquired. Most notably, we decided it was necessary to develop a custom-made response pedestal to acquire responses.

This chapter, then, focuses on the tests that were developed for computer administration and the constructs they were designed to measure. We developed tests to measure three cognitive ability constructs: Reaction Time (or Processing Efficiency), Perceptual Speed and Accuracy, and Memory, as well as three psychomotor constructs: Precision/Steadiness, Multilimb Coordination, and Movement Judgment. All but two tests were developed in time for the Fort Lewis pilot test. These two tests were included in the field test at Fort Knox (they were Number Memory and Cannon Shoot, intended as measures of the Memory and Movement Judgment constructs, respectively).

We turn now to the discussion of the development of the tests and the results of the pilot test at Fort Lewis. (Chapter 6 presents the analysis of the Fort Knox field test data).

Test Development

In our discussion of constructs, we first provide the definition and rationale for including each. Following this, the source or model used to develop each test is described, along with changes or modifications made prior to the Fort Lewis pilot test, if any. Results from the Fort Lewis pilot test are then described in detail. For example, we describe parameters used to develop test items and results from analyses of parameter data. Further, test characteristics, such as time required to read instructions and to complete the test, and test score information are provided along with recommended scoring procedures. For each test, we also highlight correlations with other computer measures and with cognitive paper-and-pencil measures. Finally, modifications or test revisions made on the basis of Fort Lewis pilot data are described.

We conclude this chapter by summarizing computer test results obtained from the Fort Lewis pilot test.

Before describing the tests designed to measure target constructs, we briefly describe a critical piece of equipment designed especially for pilot administrations of the computerized tests in the Pilot Trial Battery.

Development of Response Pedestal

The microprocessor selected for use, the COMPAQ, contains a standard keyboard. As reported in Chapter 1 and mentioned above, in early tryouts of the computer battery subjects were asked to make their responses on this keyboard. From these preliminary administrations, we determined that the keyboard may provide an unfair advantage to subjects with typing or data entry experience. Furthermore, use of a standard keyboard did not provide adequate experimental control during the testing process. Therefore, a separate response pedestal was designed and built.

This response pedestal is depicted in Figure 5.1. The pedestal is approximately 21 inches from side to side and 10 inches from front to back. Note that it contains two joy sticks (one for left-handed subjects and one for right-handed subjects), "HORIZONTAL" and "VERTICAL" controls, a dial for entering demographic data such as age and social security number, two red buttons, three response buttons--blue, yellow, and white--and four green "home" buttons. (One of the "home" buttons is not visible in the diagram; it is located on the left side of the pedestal.) The "SELECTOR" control was not used by the examinee to make responses, but was necessary to properly connect the appropriate controls to the computer for each test.

The "home" buttons play a key role in capturing subjects' reaction time scores. They control the onset of each test item or trial when reaction time is being measured. To begin a trial, the subject must place his/her hands on the four green buttons. After the stimulus appears on the screen and the subject has determined the correct response, he/she must remove his/her preferred hand from the "home" buttons and press the correct response button.

The procedure involving the "home" buttons serves two purposes. First, control is added over the location of the subjects' hands while the stimulus item is presented. In this way, hand movement distance is the same for all subjects and variation in reaction time due to position of subjects' hands is reduced to nearly zero.

Second, procedures involving these buttons are designed to assess two theoretically important components of reaction time measures--decision time and movement time. Decision time includes the period between stimulus

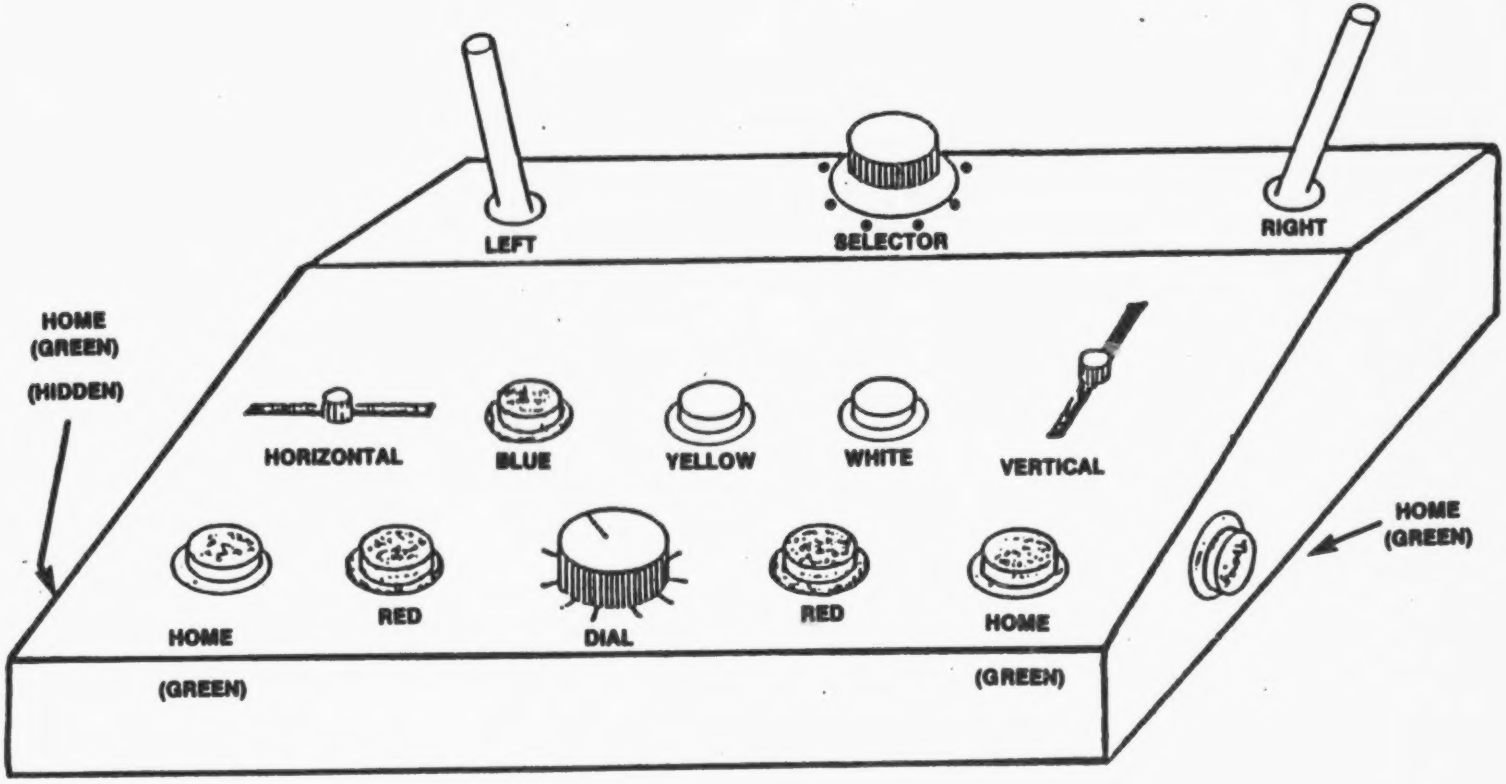

LEFT
SELECTOR
RIGHT
HOME
(GREEN)
(HIDDEN)
HORIZONTAL
BLUE
YELLOW
WHITE
VERTICAL
HOME
(GREEN)
HOME
(GREEN)
RED
DIAL
RED
HOME
(GREEN)

Figure 5.1. Response pedestal for computerized tests.

onset and the point at which the subject removes his/her hands to make a response. This interval reflects the time required to process the information to determine the correct response. Movement time involves the period between removing one's hands from the "home" buttons and striking a response key. The "home" buttons on the response pedestal, then, are designed to investigate the two theoretically independent components of reaction time. Results from an investigation of these measures appear throughout the following sections.

For each test described, we provide a schematic diagram depicting the important components of each test. A key to these schematic diagrams is provided in Figure 5.2. As noted on the key, the diagram is used to identify test components such as delay periods, operations such as decision time or movement time, and responses recorded such as correct or incorrect response, reaction time, or distance measures. These diagrams are designed to provide a more graphic picture of the activities involved in each test.

▭ = Physical operation performed by the subject

◇ = Cognitive operation performed by the subject

⬭ = Computer presentation

DP = Delay Period

DT = Decision Time

MT = Movement Time

RT = Total Reaction Time

ISD = Interstimulus Delay

R/L/B = Response Hand recorded--right, left, or both

C/I = Correct or Incorrect Response recorded

S1 = First Stimulus

S2 = Second Stimulus

d = Distance from crosshairs to the center of the target

Figure 5.2. Key to flow diagrams of computer-administered tests.

REACTION TIME (PROCESSING EFFICIENCY)

This construct involves speed of reaction to stimuli--that is, the speed with which a person perceives the stimulus independent of any time taken by the motor response component of the classic reaction time measures. According to our definition of this construct, which is an indicator of processing efficiency, it includes both simple and choice reaction time (RT).

Simple Reaction Time: Reaction Time Test 1

Test Description. The basic paradigm for this task stems from Jensen's research involving the relationship between reaction time and mental ability (Jensen, 1982). As part of this research program, Jensen designed two procedural paradigms to obtain independent measures of decision time and movement time. According to current theory, these are two independent components of reaction time. Procedures for capturing these reaction time measures are described below.

At the computer console, the subject is instructed to place his/her hands on the green "home" buttons in the ready position. When the subject is in the ready position, the first item is presented. On the computer screen, a small box appears. After a delay period (ranging from 1.5 to 3.0 seconds) the word yellow appears in the box. At this point, the subject must remove his/her preferred hand from the "home" buttons to strike the yellow key on the testing panel. The subject must then return his/her hands to the ready position to receive the next item. Figure 5.3 contains a schematic depiction of the simple reaction time task.

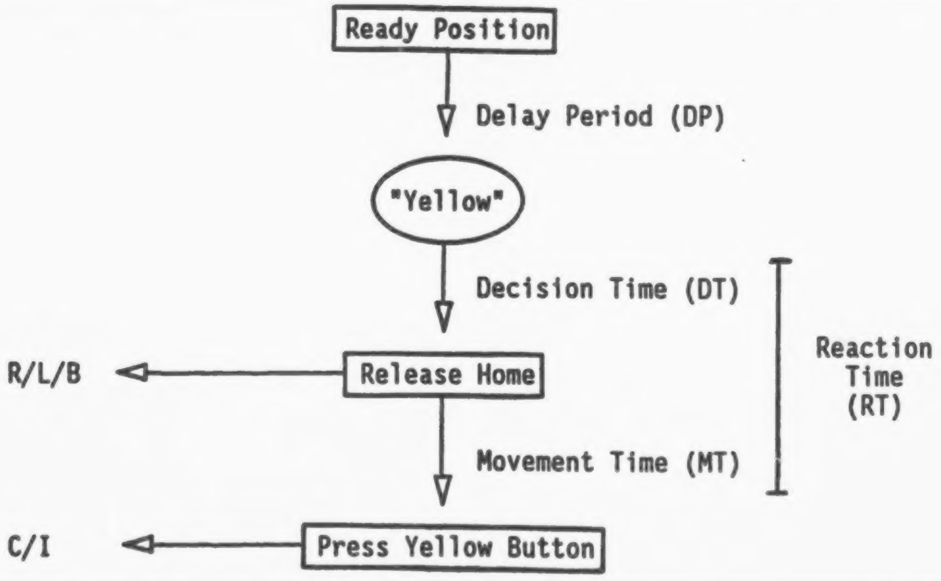


Figure 5.3. Reaction Time Test 1.

This test contains 15 items. Although it is self-paced, on each item subjects are given 10 seconds to respond before the computer time-outs[1] and prepares to present the next item.

Test Characteristics. Table 5.1 contains data on the test characteristics from the Fort Lewis pilot test. Variables appearing in the upper portion of the table provide descriptive information about test performance. Note that, on the average, subjects read the test instructions in 2.5 minutes, although this ranged from about half a minute to 5 minutes. Further, subjects completed the test in an average of 1.2 minutes; this time ranged from three-quarters of a minute to over 5 minutes. Total test time ranged from 1.6 to 7.1 minutes with a mean of 3.7 minutes.

Also note that very few subjects timed-out or provided invalid responses. The maximum number of time-outs for any subject was three, the maximum number of invalid responses was one. Finally, Percent Correct values indicate nearly all subjects understood the task and performed it correctly.

Dependent Measures[2]. To identify variables of interest, we reviewed the literature in this area. (See Keyes, *A review of the relationship between reaction time and mental ability*, 1985.) Results from this review indicated that reaction time is often calculated for decision time, movement time, and total time. See Figure 5.3 for points at which these measures are obtained. In addition, intra-individual variation measures (the standard deviation of total reaction time scores) calculated for each subject appear to provide useful information. We began isolating dependent measures of interest by calculating these four variables.

When we examined reaction times for each item on this test, we discovered that these times were very high for the first few items (up to the fifth item). Observation of the subjects when they were taking the test had alerted us to this possibility. Since this was the first test administered, the subjects were still somewhat unfamiliar with the response pedestal and the general nature of taking computer-administered tests. Accordingly, we decided to view the first five items as warm-up or practice items and to include only the last ten responses in calculating mean reaction scores.

Further, because subtle events (e.g., subject stretching or effectively guessing when the next item will appear) may produce extreme reaction time scores for a single item, we decided to use trimmed mean scores for decision, movement, and total time. These trimmed scores include responses to items six through 15 with the highest and lowest reaction time values removed.

---

[1] Time-outs occur if a subject fails to respond within a specified period of time. Invalid responses occur when a subject strikes the wrong key. In both cases, the item disappears from the computer screen and, after the subject resumes the ready position, the next item appears on the screen.

[2] Dependent variables mean scores (e.g., Decision Time) on the tests. Throughout this chapter the terms "dependent variable" and "test score" can be viewed as interchangeable.

174

Table 5.1

Pilot Test Results From Fort Lewis: Reaction Time Test 1 (Simple Reaction Time) (N = 112)

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	2.51	.81	.63 - 5.01
Time to Complete Test (minutes)	1.22	.62	.79 - 5.19
Total Test Time (minutes)	3.72	.99	1.59 - 7.10
Time-Outs (number per person)	.05	.31	0 - 3
Invalid Responses (number per person)	.07	.26	0 - 1
Percent Correct	99	3	80 - 100

Dependent Measures[a]	Mean	SD	Range	Rxx[b]
Decision Time (10 items)	30.50	10.15	17.90 - 109.78	.91
Trimmed[c] Decision Time (8 items)	29.25	8.10	18.75 - 82.00	.92
SD - Decision	7.85	12.05	.92 - 118.26	.77
Movement Time (10 items)	27.35	8.98	15.50 - 91.33	.75
Trimmed Movement Time (8 items)	26.01	7.26	15.50 - 55.86	.94
SD - Movement	6.68	12.77	.75 - 121.07	.20
Total Time (10 items)	57.84	15.78	37.90 - 149.56	.90
Trimmed Total Time (10 items)	55.92	13.86	37.75 - 124.71	.94
SD - Total	11.79	16.80	1.58 - 125.85	.66

---

[a] All values reported are in hundredths of a second.

[b] Rxx = odd-even correlations, corrected to full test length using the Spearman-Brown formula.

[c] Trimmed scores are based on response to items 6-15, excluding the highest and lowest scores.

Mean values for all the above dependent measures were calculated. They appear in the lower portion of Table 5.1. Also included in this table are reliability estimates for each measure (computed using an odd-even method with a Spearman-Brown correction). For the most part, these values are quite acceptable. Reliability for trimmed mean scores appears to be slightly higher than for those mean scores including all ten items. Further, reliability estimates for the standard deviation measures are lowest for all estimates.

To identify dependent measures for inclusion in subsequent analyses, we graphed the various reaction time scores across the 15 items. That is, mean reaction time scores were plotted for decision time, movement time, and total time across the items. These graphs indicate that movement time and total reaction time yield very similar profiles (i.e., begin at a moderately high level, drop off, and then begin to stabilize). Decision time, however, provides a slightly different profile. The graph for decision time begins at a moderately high level and drops off for the first half of the items. After that, however, it becomes very unstable and no consistent trend shows.

The relationship among these measures of reaction time was further examined by computing all pairwise correlates for each item. Mean and median values of these item-by-item correlates appear in Table 5.2 for all items (15) and for the reduced set of items (10). These results indicate that a low to moderate relationship exists between movement time and decision time (r = .32 for 10 items). Movement time appears to be providing

Table 5.2

Mean Correlations Among Decision, Movement, and Total Times: Reaction Time Test 1

	Mean	SD	Median
Decision Time and Total Times			
15 items	.61	.31	.64
10 items	.50	.29	.54
Movement and Total Times			
15 items	.80	.15	.87
10 items	.77	.16	.85
Decision and Movement Times			
15 items	.36	.25	.34
10 items	.32	.25	.30

kinds of information similar to total time ($r = .77$ for 10 items). Decision time, however, provides additional information ($r = .50$ for 10 items).

On the basis of these data, we made the following decisions:

- Subjects' scores to be analyzed should include decision time, total time, and total within-person variation score (an individual's standard deviation computed with the total score).

- For reaction time measures, the trimming procedure would be used in computing decision and total mean reaction times.

- Percent Correct scores would be computed. Although no subjects were being omitted because of incorrect or invalid responses, this could become necessary for future samples.

- Practice effects (repeating the same measure several times in a single session) should be examined, along with test-retest effects. This was planned for the Fort Knox field test.

Correlations With Other Measures. Correlations of simple reaction times with measures derived from all computer-administered tests (which are described in the sections that follow) are provided in Table 5.3.

Note that correlations among Simple Reaction Time measures (Percent Correct omitted from this analysis) indicate that the three correlate very highly with one another (Decision with Total = .85; Total SD with Total = .67; and Decision with Total SD = .71). Decision and Total times for simple reaction time correlate moderately with their Choice Reaction Time counterparts (range .36 to .57) which are described in the next section.

Correlations of Simple Reaction Time measures with computer test dependent measures from constructs other than processing efficiency, indicate that for Decision Time the highest correlations appear with Perceptual Speed and Accuracy (PS & A) Intercept (.30), Grand Mean (.29), and Memory Intercept (.30). Total time also correlates highest with PS & A Intercept (.45). Total Standard Deviation correlates highest with Memory Intercept (.29). These correlations are about as expected since the correlated scores are all reaction times to intercepts based on reaction times for perceptual kinds of tests. (Memory involves a perceptual component even though it is primarily a measure of the Memory construct.)

Correlations of the various computer-administered measures with the cognitive paper-and-pencil measures described in Chapters 3 and 4 are shown in Table 5.4. These correlations indicate that Decision Time, Total Standard Deviation, and Percent Correct are virtually unrelated to scores on the paper-and-pencil measures. Total reaction time, however, correlates highest with the Maze (-.39), Path (-.23), and Orientation 1 (-.23) Tests. These negative correlations indicate that "better" (faster) total reaction time scores are associated with better (higher) paper-and-pencil test scores.

Finally, scores on these measures were correlated with video experi-

Table 5.3

Intercorrelations of Dependent Measures Developed From Computer-Administered Tests: Fort Lewis Pilot Test

	Reaction Time						Perceptual Speed and Accuracy			Short-Term Memory				Tracking		Target Shoot	Target Identification	
	Simple - Decision Time	Simple - Total Time	Simple - Total SD	Choice - Decision Time	Choice - Total Time	CRT-SRT Total	PSA Slope	PSA Intercept	PSA Grand Mean	Memory Slope	Intercept	% Correct	Grand Mean	Tracking 1 1-Mean	Tracking 2 2-Mean	Target Shoot Mean Distance	Target ID Mean-RT	% Correct
Reaction Time (RT)																		
Simple - Decision Time	**	.85	.71	.36	.37	-.66	-.01	.30	.16	-.03	.30	-.11	.29	.20	-.09	-.01	.17	.11
Simple - Total Time		**	.67	.36	.57	-.65	-.04	.45	.22	-.04	.32	-.20	.30	.31	.11	.04	.31	-.12
Simple - Total SD			**	.05	.10	-.69	-.01	.14	.07	-.11	.29	-.14	.22	.08	-.15	-.01	.10	.05
Choice - Decision Time				**	.78	.31	.05	.37	.27	.05	.29	-.06	.33	.15	.14	.13	.29	-.08
Choice - Total Time					**	.25	.04	.53	.36	-.02	.41	-.11	.40	.39	.33	.14	.45	-.07
CRT-SRT (Total)						**	.09	-.04	.08	.03	.00	.14	.02	.00	.17	.08	.06	.08
Perceptual Speed and Accuracy (PS & A)																		
Slope							**	-.27	.79	.06	.09	.29	.13	.06	.07	.05	.30	.27
Intercept								**	.35	.04	.44	-.43	.48	.36	.31	.19	.45	-.23
Grand Mean RT									**	.08	.37	.01	.43	.29	.27	.18	.58	.13
Short-Term Memory																		
Slope										**	-.31	-.13	.29	-.06	.02	.03	.13	-.07
Intercept											**	-.33	.82	.25	.17	.10	.45	.09
Percent Correct												**	-.41	-.14	-.24	-.25	-.15	.31
Grand Mean RT													**	.21	.19	.12	.54	.05
Tracking																		
Test 1 - Mean Distance														**	.76	.32	.46	-.10
Test 2 - Mean Distance															**	.47	.44	-.11
Target Shoot																		
Mean Distance																**	.16	-.10
Target Identification																		
Mean RT																	**	.21
Percent Correct																		**

Table 5.4

Intercorrelations of Cognitive Paper-and-Pencil Tests and Computer-Administered Tests: Fort Lewis Pilot Test

	Assembling Objects	Object Rotation	Path	Maze	Shapes	Orientation 1	Orientation 2	Orientation 3	Reasoning 1	Reasoning 2
Reaction Time (RT)										
Simple-Decision Time	-01	-03	-10	-23	-10	-05	06	01	-06	04
Simple-Total Time	-10	-15	-23	-39	-21	-23	-09	-17	-20	-14
Simple-Total SD	-01	-01	-10	-13	-07	-05	00	-03	-01	-01
Simple-Percent Correct	01	-07	17	02	04	07	-02	00	08	10
Choice-Decision Time	-09	-12	-17	-28	-21	-18	-17	-15	-12	-15
Choice-Total Time	-22	-27	-23	-47	-32	-36	-26	-29	-25	-25
Choice-Total SD	-20	-12	00	-05	-22	-17	-17	-15	-07	-07
Choice-Percent Correct	07	-10	-01	-05	-05	-08	-10	-05	08	-07
Perceptual Speed & Accuracy (PS & A)										
Slope	16	-12	11	01	-03	09	19	11	14	27
Intercept	-44	-40	-46	-57	-37	-50	-42	-44	-48	-43
Percent Correct	30	09	26	16	17	31	21	25	20	31
Grand Mean RT	-11	-35	-17	-33	-24	-22	-09	-17	-16	-01
Short-Term Memory										
Slope	04	03	-03	13	-02	-10	-04	-04	04	06
Intercept	-22	-30	-24	-40	-17	-26	-08	14	-23	-22
Percent Correct	29	17	46	34	17	31	25	29	32	28
Grand Mean RT	-20	-29	-26	-33	-18	-32	-11	-16	-21	-19
Tracking										
Test 1 Mean Distance	-27	-45	-39	-52	-41	-39	-29	-39	-38	-30
Test 2 Mean Distance	-32	-46	-43	-50	-36	-45	-38	-44	-35	-33
Target Shoot										
Mean Distance	-13	-14	-20	-23	-22	-21	-22	-18	-17	-10
Percent Correct	25	27	20	40	30	28	27	27	21	18
Target Identification										
Mean RT	-30	-46	-31	-50	-39	-48	-32	-43	-42	-32
Percent Correct	27	17	24	29	17	11	16	11	26	19

NOTE: Decimals have been omitted.

ence.[3] Mean Decision Trimmed and Mean Total Trimmed correlate near zero with this variable. Total Standard Deviation correlates .19 and Percent Correct correlates -.20 with this measure.

Modifications for Fort Knox Field Test. The Reaction Time Test 1 administered in the Fort Lewis pilot test remained the same for the Fort Knox field test.

Choice Reaction Time: Reaction Time Test 2

Test Description. Reaction time for two response alternatives (choice reaction time, CRT) is obtained in virtually the same manner as for a single response (simple reaction time, SRT). The major difference is in stimulus presentation. Rather than the same stimulus, YELLOW, being presented, the stimulus varies; that is, subjects may see the term BLUE or WHITE on the computer screen. When one of these terms appears, the subject is instructed to move his/her preferred hand from the "home" keys to strike the key that corresponds with the term appearing on the screen (BLUE or WHITE). See Figure 5.4 for a schematic depiction of the test.

This measure contains 15 items, with seven requiring responses on the WHITE key and eight requiring responses on the BLUE key. Although the test is self-paced, the computer is programmed to allow 9 seconds for a response before going on to the next item. Data for all 15 items were included in the analysis of the data from the Fort Lewis pilot test. The subjects had become familiar enough with the response pedestal that it was not thought necessary to treat any items as "warm-ups."

Test Characteristics. Table 5.5 provides data describing this test as it was given in the Fort Lewis pilot test. Note that subjects were reading the instructions more quickly than they were for simple reaction time (1.01 and 2.51 minutes, respectively) and were also finishing the test more quickly (1.95 and 3.72 minutes, respectively).

Data on whether subjects used the same or different hands to respond to all items indicate that 23 percent of the subjects (N=26) consistently used the same hand. The remainder (77% or N=86) switched from hand to hand at least once to respond.

We also examined reaction time differences in responding to the BLUE and WHITE keys. These results indicate that, on the average, subjects responded a little faster to the WHITE versus the BLUE key (64.92 versus 69.12 hundredths of a second).

Dependent Measures. In the description of simple reaction time, we provided a rationale for the measures selected to score subjects' responses. These same measures were also selected to score responses on choice reaction time. Mean values along with standard deviations, ranges, and reliability estimates are provided in Table 5.5. Note that for this

---

[3] Subjects were asked to rate, on a five-point scale, their degree of experience with video game playing, prior to completing the computer tests. (A rating of 1 indicated no experience with video games; 5 indicated much experience.)

181

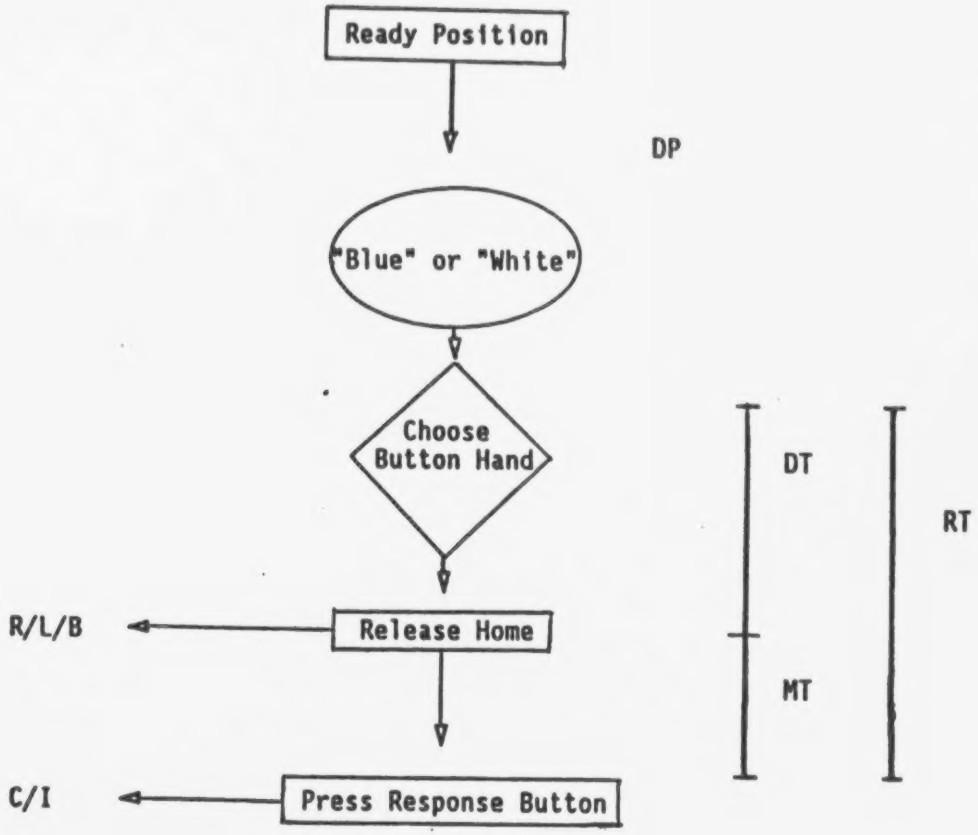

Ready Position
DP
"Blue" or "White"
Choose Button Hand
DT
RT
R/L/B
Release Home
MT
C/I
Press Response Button

Figure 5.4. Reaction Time Test 2.

Table 5.5

Pilot Test Results from Fort Lewis: Reaction Time Test 2 (Choice Reaction Time) (N = 112)

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	1.01	.36	.45 - 2.37
Time to Complete Test (minutes)	.95	.13	.80 - 1.59
Total Test Time (minutes)	1.95	.40	1.37 - 3.20
Time-Outs (number per person)	0	0	0 - 1
Invalid Responses (number per person)	.17	.10	0 - 1

Dependent Measures	Mean	SD	Range	Rxx[a]
Mean Decision Time[b]	36.78	7.76	18.75 - 78.29	.94
Mean Total Time[b]	65.98	10.38	37.75 - 117.29	.91
SD - Total Time[b]	8.92	3.75	1.09 - 60.07	.10
Percent Correct	99	3	90 - 100	-.16

Choice RT Minus Simple RT	Mean	SD	Range	Rxx[a]
Decision Time[b]	7.68	8.79	-43.70 - 33.99	.86
Total Time[b]	10.37	11.15	-44.92 - 38.71	.79

---

[a] Rxx = odd-even correlations corrected with the Spearman-Brown formula.

[b] Values reported are in hundredths of a second. Statistics are based on analysis of all 15 items of the test.

measure, only the two reaction time scores provide reliable information.

Another measure we looked at is the difference between mean choice reaction time scores and simple reaction time scores--a value that is intended to capture a speed-of-processing component. The typical choice reaction time paradigm includes two, four, and eight response alternatives, and processing efficiency is computed by regressing mean reaction time score against the number of response alternatives (i.e., one, two, four, and eight). The slope of this regression equation is interpreted as the processing speed, or the time required to process additional information. Because our testing pedestal does not allow for four or eight response choices, we cannot calculate this value. Instead, we used a score showing the difference between choice and simple reaction times. Note that reliability estimates suggest these values are internally consistent.

Correlations With Other Measures. Correlations with measures derived from other computer-administered tests are reported in Table 5.3. These values indicate that choice decision and choice total times are highly correlated ($r = .78$). (Standard deviation total and percent correct were omitted from these analyses due to low reliability.) Choice decision and choice total times correlate moderately with their simple reaction time counterparts. Also note that the experimental variable, Choice Total Time minus Simple Total Time, correlates highly with Simple Reaction Time measures, but only moderately with Choice Reaction Time measures.

Choice Decision and Choice Total yield fairly similar correlation patterns with scores from other computer tests. These measures correlate highest with PS & A Intercept ($r = .37$ and $r = .53$, respectively), Target Identification Mean RT ($r = .29$ and $r = .45$), and Memory Intercept ($r = .29$ and $r = .41$) and Grand Mean ($r = .33$ and $r = .40$). In addition, choice total yields moderate correlations with Tracking 1 Mean ($r = .39$), Tracking 2 Mean ($r = .33$), and PS & A Grand Mean ($r = .36$). Again, just as for Simple Reaction Time, these correlations show an association between reaction times for the perceptual tasks--except for the moderate correlations with Tracking 1 and 2, which are somewhat unexpected, but may indicate association based on movement speed.

Correlations of choice reaction time measures with cognitive paper-and-pencil measures appear in Table 5.4. These data indicate that choice decision and total time correlate highest with the Maze Test ($r = -.28$ and $-.47$, respectively). Total time, in fact, yields moderate correlations across all paper-and-pencil cognitive measures. As noted before, these negative correlations actually indicate that "better" scores are associated since lower scores on reaction time indicate better performance and higher scores on the paper-and-pencil tests indicate better performance.

Modifications for Fort Knox Field Test. No changes were made to this test for the Fort Knox field test.

SHORT-TERM MEMORY

This construct is defined as the rate at which one observes, searches, and recalls information contained in short-term memory.

Memory Search Test

The marker used for this test is a short-term memory search task introduced by S. Sternberg (1966, 1969). In this test, the subject is presented with a set of one to five familiar items (e.g., letters); these are withdrawn and then the subject is presented with a probe item. The subject is to indicate, as rapidly and as accurately as possible, whether or not the probe was contained in the original set of items, now held in short-term memory. Generally, mean reaction time is regressed against the number of objects in the item or stimulus set. The slope of this function can be interpreted as the average increase in reaction time with an increase of one object in the memory set, or the rate at which one can access information in short-term memory.

Test Description. The measure developed for computer-administered testing is very similar to that designed by Sternberg. At the computer console, the subject is instructed to place his/her hands on the green home buttons. The first stimulus set then appears on the screen. A stimulus contains one, two, three, four, or five objects (letters). Following a .5- or 1-second display period, the stimulus set disappears and, after a delay, the probe item appears. Presentation of the probe item is delayed by either 2.5 or 3 seconds. When the probe appears, the subject must decide whether or not it appeared in the stimulus set. If the item was present in the stimulus set, the subject removes his/her hands from the home buttons and strikes the white key. If the probe item was not present, the subject strikes the blue key. (See Figure 5.5 for schematic depiction of the memory search task.) Fifty items were included on this test for the Fort Lewis administration.

Parameters of interest include, first, stimulus set length, or number of letters in the stimulus set. Values for this parameter range from one to five. The second parameter, observation period and probe delay period, includes two levels. The first is described as long observation and short probe delay; time periods are 1 second and 2.5 seconds, respectively. The second level, short observation and long probe delay, includes periods of .5 second and 3 seconds, respectively. The final parameter, probe status, indicates that the probe is either in the stimulus set or not in the stimulus set. These parameters will be discussed in more detail below.

Test Characteristics. Table 5.6 provides descriptive information for the Memory Search Test from the pilot test at Fort Lewis. These data indicate that subjects, on the average, read the test instructions in 3 minutes (range, 1.6 - 5.8) and completed the test in 9 minutes (range, 8.4 - 11.7). Thus, total testing time for the average subject is 12 minutes (range, 10.4 - 17.5). Further, subjects allowed very few timeouts (mean = .17, SD = .80) and provided about five invalid responses (range 0 - 28). Over all, total percent correct is 90. However, the range of Percent Correct values, 44 to 100, indicates that at least one subject was performing at a lower than chance level.

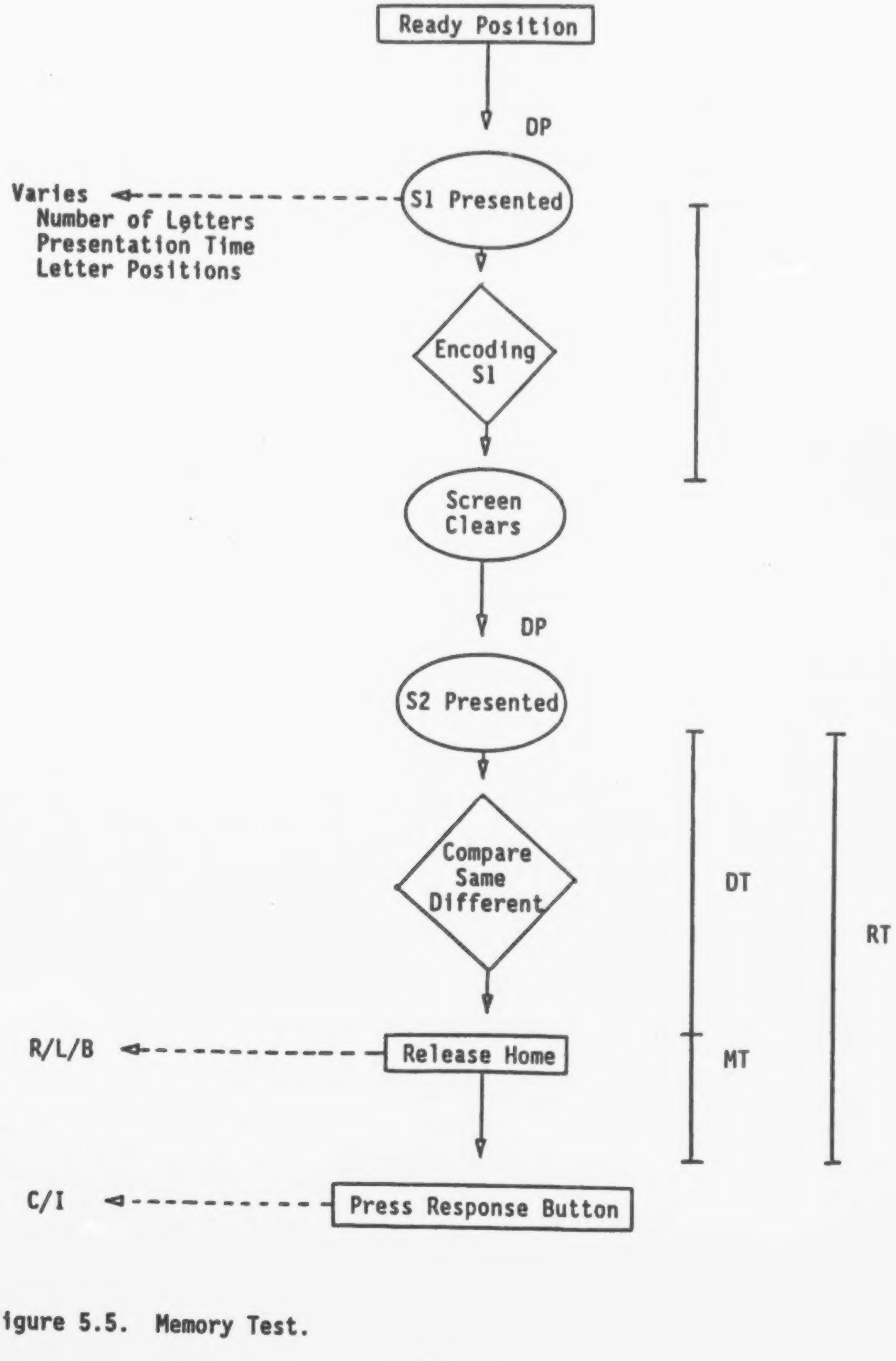

Ready Position
DP
Varies
Number of Letters
Presentation Time
Letter Positions
S1 Presented
Encoding S1
Screen Clears
DP
S2 Presented
Compare Same Different
DT
RT
R/L/B
Release Home
MT
C/I
Press Response Button

Figure 5.5. Memory Test.

Table 5.6

Pilot Test Results From Fort Lewis: Memory Search
(N =112)

Test Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	3.06	.76	1.64 - 5.81
Time to Complete Test (minutes)	9.00	.54	8.37 - 11.71
Total Test Time (minutes)	12.07	1.06	10.43 - 17.52
Time-Outs (number per person)	.17	.80	0 - 8
Invalid Responses (number per person)	4.86	4.72	0 - 28

Dependent Measures[a]	Mean	SD	Range	Rxx[b]
Slope[c]	7.19	6.14	-12.70 - 41.53	.54
Intercept[c]	97.53	30.28	44.91 - 230.97	.84
Grand Mean[c]	119.05	29.84	67.71 - 262.35	.88
Percent Correct	89	10	44 - 100	.95

---

[a] See text for explanation of these measures.

[b] Rxx = odd-even correlation corrected with the Spearman-Brown formula.

[c] Values reported are in hundredths of a second. Statistics are based on an analysis of items answered correctly. (There were 50 items on the test.)

Dependent Measures. For this test, mean values for decision time, movement time, and total time were computed and then plotted against item length, defined as the number of letters in the stimulus set. These plots indicated that decision and total time produce very similar profiles, whereas movement time results in a nearly flat profile. Since decision time and total time yield similar information and movement time appears to serve as a constant, we could have used either decision or total reaction time to compute scores on this measure. We elected to use total reaction time.

Subjects receive scores on the following measures:

- Slope and Intercept. These values are obtained by regressing mean total reaction time (correct responses only) against item length. In terms of processing efficiency, Slope represents the average increase in reaction time with an increase of one object in the stimulus set. Thus, the lower the value, the faster the access. Intercept represents all other processes not involved in memory search, such as encoding the probe, determining whether or not a match has been found, and executing the response.

- Percent Correct. This value is used to screen subjects completing the test. For example, recall that in Table 5.6 we indicated that one subject correctly answered 44 percent of the items. Computing the above scores (e.g., Slope and Intercept) for this subject would result in meaningless information. Thus, Percent Correct scores are used to identify subjects performing at very low levels, thereby precluding computation of the above scores.

- Grand Mean. This value is calculated by first computing the mean reaction time (correct responses only) for each level of stimulus set length (i.e., one to five). The mean of these means is then computed.

Table 5.6 contains the mean, standard deviation, range, and reliability estimates for each of the dependent measures. Note that these values indicate that all measures except the Slope yield fairly high internal consistency values.

Correlations With Other Measures. The four dependent measures computed for the Short-Term Memory Test were correlated with scores generated from the other computer-administered tests of the battery and with scores on the cognitive paper-and-pencil tests (Tables 5.3 and 5.4, respectively). Results for these four dependent measures varied, and are discussed separately.

Short-Term Memory Slope yielded correlations ranging from -.31 to .29 with other computer measures. Lowest values were with Choice Reaction Time Total (*r* = -.02) and Target Tracking 2 (*r* = .02), while highest values were with Memory Intercept (*r* = -.31) and Grand Mean (*r* = .29). Dependent measures from other computer tests correlating moderately with Memory Slope include Simple Reaction Time Total SD (*r* = -.11) and Target Identification Mean Reaction Time (*r* = .13). When correlated with cognitive paper-and-pencil tests, Short-Term Memory Slope yielded generally low relationships. The highest correlation was .13 with the Maze Test.

Short-Term Memory Intercept correlated highest with the Memory Grand Mean (*r* = .82), Target Identification Mean Reaction Time (*r* = .45), Perceptual Speed and Accuracy Intercept (*r* = .44), and Choice Reaction Time Total (*r* = .41). Low relationships were found with the difference between choice and simple reaction times (*r* = .00), Perceptual Speed and Accuracy Slope (*r* = .09), Target Shoot Mean Distance (*r* = .10), and Target Identification Percent Correct (*r* = .09). With the cognitive paper-and-pencil measures, Memory Intercept showed generally moderate relationships, for example, with Maze (*r* = -.40), Object Rotation (*r* = -.30), and Orientation 1 (*r* = -.26).

Short-Term Memory Percent Correct correlated most strongly with Perceptual Speed and Accuracy Intercept (*r* = -.43), and with other measures on the Memory Test (*r* = -.33 with Intercept, *r* = -.41 with Grand Mean). Weak correlations were found between Short Term Memory Percent Correct and Choice Reaction Decision Time (*r* = -.06) and Perceptual Speed and Accuracy Grand Mean (*r* = .01). It correlated fairly highly with Path (*r* = .46) and moderately with several other cognitive written tests, while the lowest coefficients were with Object Rotation and Shapes (*r* = .17 for both).

Finally, the last dependent measure of the Short-Term Memory Test was the Grand Mean Reaction Time (for correct responses only). This correlated most highly with the computer measures of Mean Reaction Time on Target Identification (*r* = .54) and the Perceptual Speed and Accuracy Intercept (*r* = .48), as well as the Short-Term Memory Intercept (*r* = .82). Lowest correlations were found with the difference between choice and simple reaction time (*r* = .02) and with the Target Identification Percent Correct (*r* = .05). Strongest relationships with the cognitive paper-and-pencil tests were found between the Short-Term Memory Grand Mean and Maze (*r* = -.33) and Orientation 1 (*r* = -.32). Lowest were with Orientation 2 and 3 (*r* -.11 and -.16, respectively).

To sum up these correlations, the Grand Mean RT and Intercept for memory show highly similar patterns of correlations with other computer-administered tests and with cognitive paper-and-pencil tests. Both measures are moderately correlated with Reaction Time scores and Intercept scores on other computer-administered tests, and have low to moderate correlations with paper-and-pencil test scores. The Slope score for memory shows low correlations with scores on almost all other measures. The Percent Correct score for memory shows low to moderate negative correlations with Reaction Time and Intercept scores on other computer-administered measures, and moderate correlations with scores on cognitive paper-and-pencil tests. These patterns of correlations are about as expected and seem to indicate that the Memory Test scores contribute some fairly unique variance to the PTB.

Modifications for Fort Knox Field Test. Results from an analysis of variance (ANOVA) conducted for the Fort Lewis pilot test data were used to modify this test for the Fort Knox field test. As noted earlier, the three parameters were stimulus set length, observation period/probe delay, and probe status. Total reaction time served as the dependent variable for this measure. A three-way ANOVA, 5 (stimulus set length) x 2 (observation period/probe delay) x 2 (probe status), was performed.

These data indicated that the two levels of observation period and probe delay yielded no significant differences in reaction time ($F = .27$; $p<.60$). For stimulus set length, levels one to five, mean reaction time scores differed significantly ($F = 84.35$; $p<.001$). This information confirms results reported in the literature; that is, reaction time increases as stimulus set length increases. Finally, for probe status, in or not in, mean reaction time scores also differed significantly ($F = 74.24$; $p<.001$). These values indicate that subjects, on the average, require more time to determine that a probe is not in the set than to determine that the probe is contained in the set. Results also indicated a significant interaction between stimulus length and probe status ($F = 7.46$; $p<.001$).

This information was used to modify the Memory Search Test. For example, stimulus set length had yielded significant mean reaction time score differences for the five levels; mean reaction time for levels two and four, however, differed little from levels three and five, respectively. Therefore, items containing stimulus sets with two and four letters were deleted from the test file.

Although the observation period/probe delay parameter produced nonsignificant results, we concluded that different values for probe delay may provide additional information about processing and memory. For example, in literature in this area researchers suggest that subjects begin with a visual memory of the stimulus objects, which begins to decay after a very brief period, .5 second. To retain a memory of the object set, subjects shift to an acoustic memory; that is, subjects rehearse the sounds of the object set and recall its contents acoustically (Thorson, Hochhaus, & Stanners, 1976). Therefore, we changed the two probe delay periods to .5 seconds and 2.5 seconds. These periods are designed to assess the two hypothesized types of short-term memory--visual and acoustic.

Finally, consideration of the probe status parameter led us to modify one-half of the items in the test to include unusual or unfamiliar objects--symbols, rather than letters. In part, rationale for using letters or digits in a problem involves using overlearned stimuli so that novelty of the stimulus does not affect processing of the material. We elected, however, to add a measure of processing and recalling unusual material, primarily because Army recruits do encounter and are required to recall stimuli that are novel to them, especially during their initial training. Consequently, one-half of the revised test items ask subjects to observe and recall unfamiliar symbols rather than letters.

The test then, as modified, contained 48 items--one half consisting of letters and the other half of symbols. Within each item type, three levels of stimulus length are included. That is, for items with letter stimulus sets, there are eight items with a single letter, eight with three, and eight with five letters; the same is done for items containing symbols. Within each of the stimulus length sets, four items include a .5-second probe delay and four contain a 2.5-second probe delay period. Across all items (N = 48), probe status is equally mixed between "in" and "not in" the stimulus set. With the test so constructed, the effects of stimulus type, stimulus set length, probe delay period, and probe status can be examined.

PERCEPTUAL SPEED AND ACCURACY

The perceptual speed and accuracy (PS & A) construct involves the ability to perceive visual information quickly and accurately and to perform simple processing tasks with the stimulus (e.g., make comparisons). This requires the ability to make rapid scanning movements without being distracted by irrelevant visual stimuli, and measures memory, working speed, and sometimes eye-hand coordination.

Perceptual Speed and Accuracy Test

Measures used as markers for the development of the computer-administered Perceptual Speed and Accuracy Test included such tests as the Employee Aptitude Survey (EAS-4) Visual Speed and Accuracy, and the ASVAB Coding Speed test and the Tables and Graphs test. The EAS-4 involves the ability to quickly and accurately compare numbers and determine whether they are the same or different, whereas ASVAB Coding Speed measures memory, eye-hand coordination, and working speed. The Tables and Graphs test requires the ability to obtain information quickly and accurately from material presented in tabular form.

Test Description. The computer-administered Perceptual Speed and Accuracy Test requires the subject to make a rapid comparison of two visual stimuli presented simultaneously and determine whether they are the same or different. Five different "types" of stimuli are presented: alpha, numeric, symbolic, mixed, and words. Within the alpha, numeric, symbolic, and mixed stimuli, the character length of the stimulus is varied; four different levels of stimulus length or "digit" are present--two-digit, five-digit, seven-digit, and nine-digit. Four items are included in each "type" x "digit" cell; for example, four items are two-digit alphas (e.g., XA). In its original form this test had:

16	two-digit items
16	five-digit items
16	seven-digit items
16	nine-digit items
16	word items
80	total items

Same and different responses were balanced in every cell except one; the four two-digit numeric items were accidentally constructed to require all "same" responses. Some example items are shown below:

1.	96293	96298	(Numeric five-digit)
2.	$+/^{0*}<>^{2}$	$+/^{0*}<>^{2}$	(Symbolic seven-digit)
3.	James Braun	James Brown	(Words)

Reaction times were expected to increase with the number of digits included in the stimulus. The rationale behind including various types of stimuli was simply that various types of stimuli are often encountered in military positions.

The subject is instructed to hold the home keys down to begin each item, release the home keys upon deciding whether the stimuli are the same or different, and press the white button if the stimuli are the same or the blue button if the stimuli are different (see Figure 5.6).

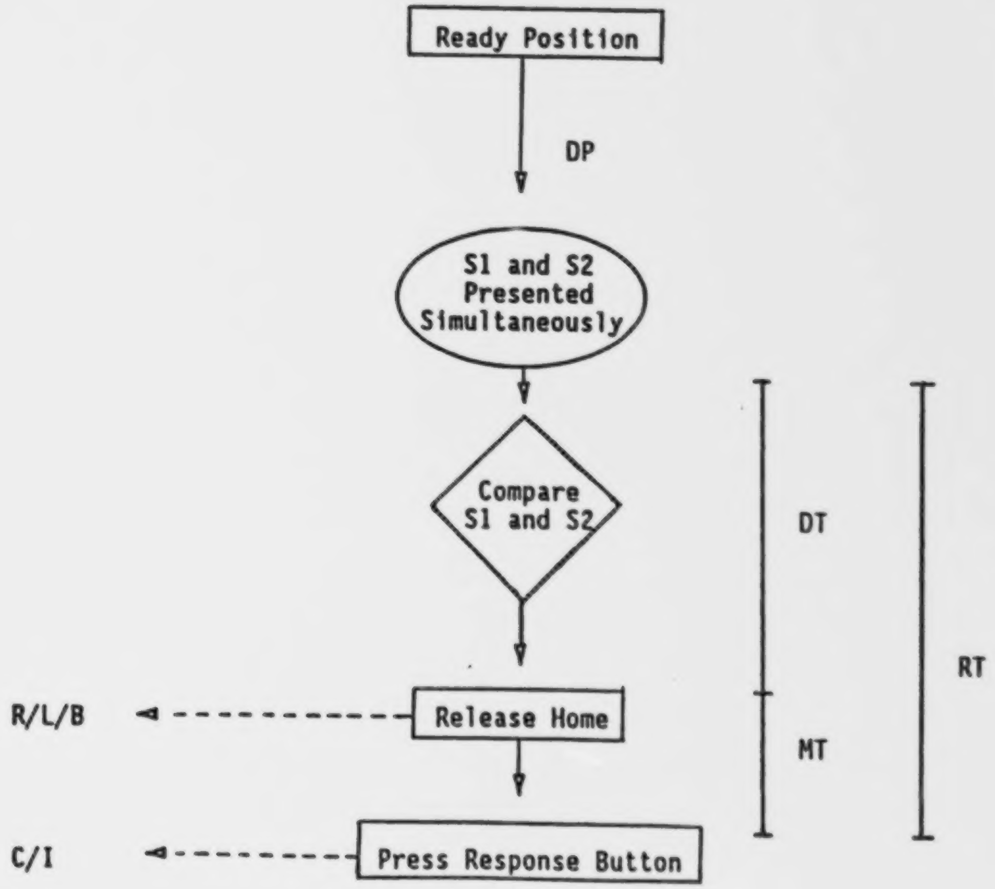

Ready Position
DP
S1 and S2 Presented Simultaneously
Compare S1 and S2
DT
RT
R/L/B
Release Home
MT
C/I
Press Response Button

Figure 5.6. Perceptual Speed and Accuracy Test.

Test Characteristics. The computerized Perceptual Speed and Accuracy Test was administered to 112 individuals in the pilot test at Fort Lewis. Some of the overall test characteristics are shown in Table 5.7.

Table 5.7

Pilot Test Results From Fort Lewis: Overall Characteristics of Perceptual Speed and Accuracy Test (N = 112)

	Mean	SD	Range
Time Spent on Instructions (minutes)	2.36	.59	1.37 - 4.30
Time Spent on Test Portion (minutes)	7.82	1.04	5.82 - 12.41
Total Testing Time (minutes)	10.18	1.37	7.45 - 14.88
Time-Outs (number per person)	9.57	6.17	0 - 35
Invalid Responses (number per person)	.94	1.20	0 - 6

The average total testing time was just over 10 minutes (range = 7.4 to 14.9 minutes). Subjects were given 7 seconds to respond to each item. There were more time-outs on this test (mean = 9.6) than on the previously described tests. On the other hand, there were fewer invalid responses than on Short-Term Memory (mean = .94 for Perceptual Speed and Accuracy vs. 4.86 for Short-Term Memory).

Dependent Measures. The measures obtained were: response hand, percent correct, total reaction time, decision time, movement time, time for instructions, and total test time. The variables to be used for scoring purposes or dependent measures were determined through results of ANOVAs on total reaction times. The resulting variables include:

The grand mean of the mean reaction times for each digit level for correct responses only.

The mean total reaction time of "word" items for correct responses only.

The slope and intercept for the regression of mean total reaction time on digits for correct responses (i.e., intercept and the change in total reaction time per unit change in stimulus length).

The grand mean of the mean reaction times for the four "non-word" digit levels and the "word" items.

The percent of all items answered correctly.

The rationale behind the selection of these variables will be provided in the discussion of the ANOVA results.

Two two-way ANOVAs were performed on reaction times for correct responses. The first was a Type (4 levels) x Digit (4 levels) ANOVA of total reaction times. The results showed significant main effects for Type [F (3,333) = 11.99, $p<.001$], Digits [F (3,333) = 871.46, $p<.001$], and their interaction [F (9,999) = 44.14, $p<.001$] (see Figure 5.7).

The second ANOVA conducted was on movement times. Pure movement time should be a constant when response hands are balanced. The results suggested that subjects were still making their decision about the stimuli after releasing the home keys (see Figure 5.8). That is, the movement time ANOVA for Type X Digits yielded a significant main effect for Digits [F (3,333) = 19.94, $p<.001$]. The interaction of Digits and Type was also significant [F (9,999) = 7.22, $p<.001$].

The implications of these results are:

Scores should be formed on total reaction times (for correct responses) instead of decision times because subjects appear to continue making a decision after releasing the home keys. Thus, use of decision time would not include time that subjects were using to process items.

Means should be computed separately for each set of items with a particular digit level (i.e., two, five, seven, and nine). Number of digits had a greater effect on mean reaction time than did type of stimuli. Since only correct response reaction times are being used, subjects could raise their scores on a pooled reaction time by simply not responding to the nine-digit items. Thus, the mean reaction times to correct responses for each digit level should be equally weighted. The grand mean of the mean reaction times for each digit level was computed.

The nine-digit symbolic items were probably too easy. Mean reaction times for the nine-digit symbolic items were substantially less than those for the other nine-digit items. Further inspection of the items showed that some were probably being processed in "chunks" because symbols were grouped (e.g., <<++++*//).

Total reaction times for correct responses could be regressed on digit. Intercepts and slopes could be computed for individuals by means of a repeated measures regression (i.e., the trend appeared to be linear).

As a whole, the scores on the computerized Perceptual Speed and Accuracy Test were quite reliable (see Table 5.8). Reliability coefficients ranged from .85 for the Intercept of the regression of total reaction time on digits to .97 for the Grand Mean of the mean reaction times for the four non-words categories and for all categories.

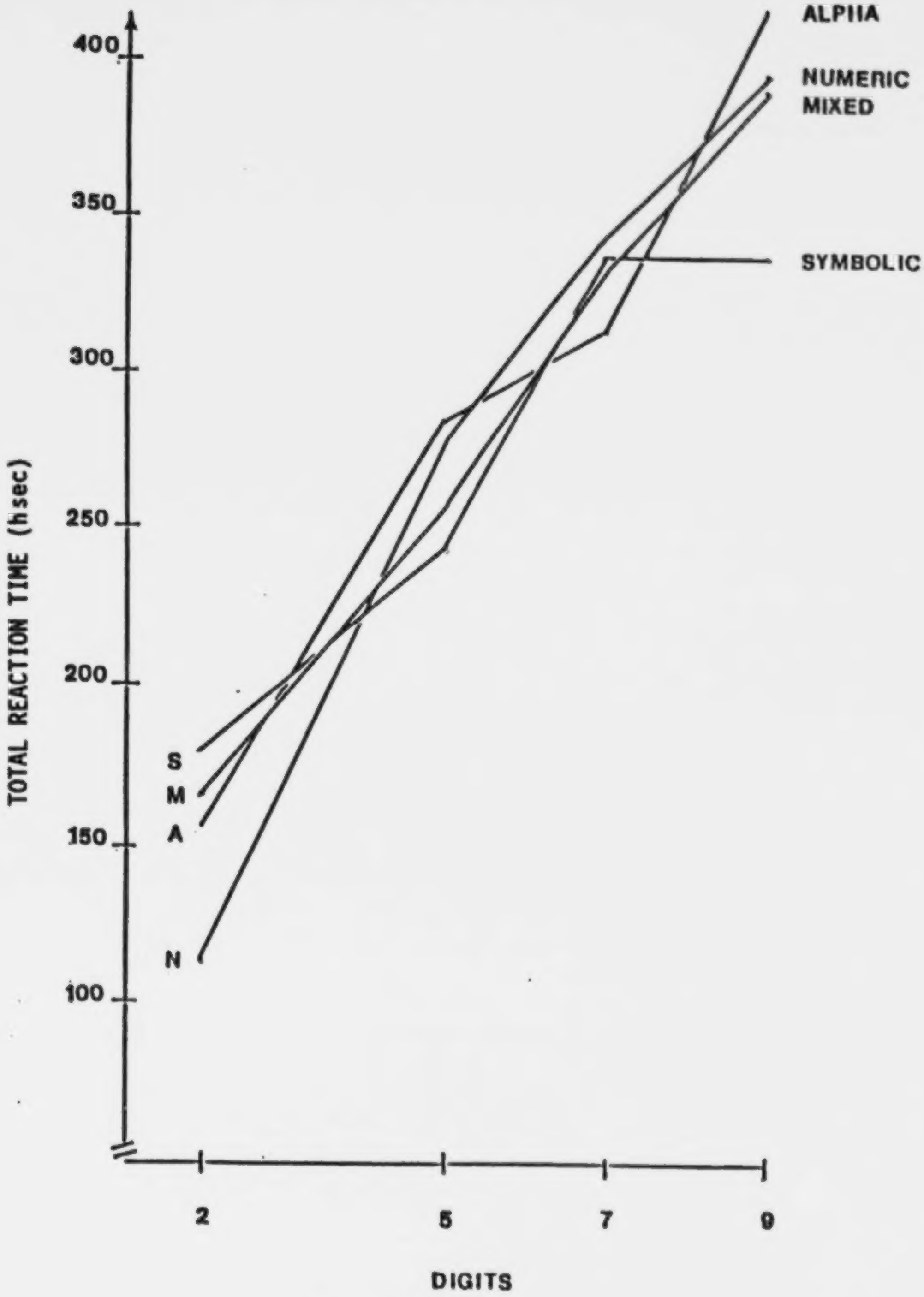


Figure 5.7. Type x Digit analysis of variance on Total Reaction Time.

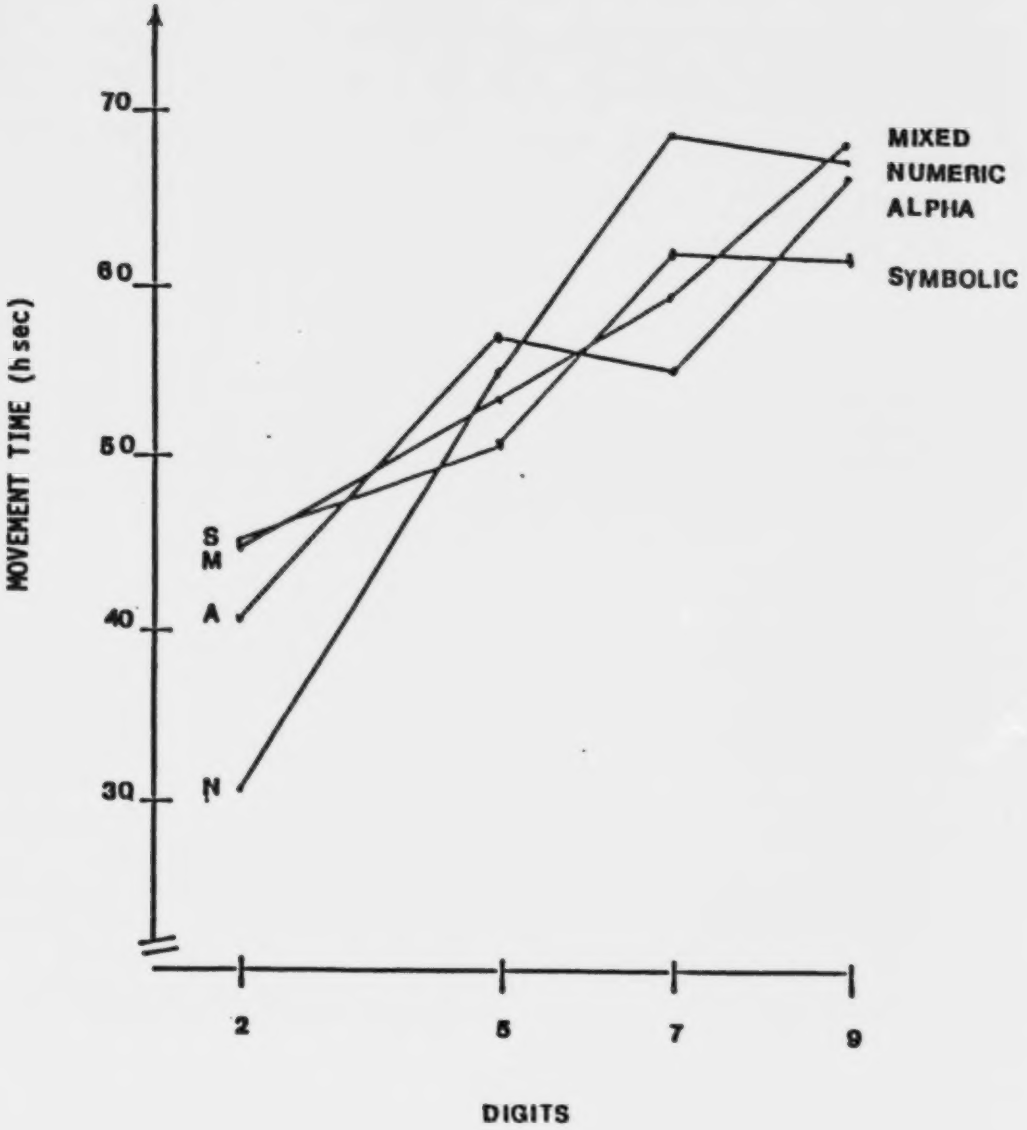


Figure 5.8. Type x Digit analysis of variance on Movement Time.

Table 5.8

Pilot Test Results From Fort Lewis: Dependent Measure Scores From Perceptual Speed and Accuracy Test (N = 112)

Score[a]	Mean	SD	Range	Rxx[b]
Grand Mean of Mean Reaction Times for Non-word Items	279.99	57.97	85.67 - 386.49	.97
Mean Reaction Time for Word Items	351.74	68.39	198.64 - 518.64	.91
Grand Mean of Mean Reaction Times for Word and Non-word Items	294.22	57.13	109.34 - 412.75	.97
Intercept	89.37	36.48	12.99 - 210.34	.85
Slope	33.14	9.78	-.75 - 52.11	.89
Percent Correct	86.90	8.00	56.3 - 100	

[a] Reaction Time values are in hundredths of a second and are based on analysis of items answered correctly. (There were 80 items on the test.)

[b] Split-half (odd-even) reliability estimates, Spearman-Brown corrected.

Interrelationships Among Perceptual Speed and Accuracy Scores. Ideally, efficient performance on the Perceptual Speed and Accuracy Test would produce: a low intercept, a low slope, and high accuracy, combined with a fast grand mean reaction time score. Data analyzed from the Fort Lewis testing suggests that this relationship may occur infrequently. As shown in Table 5.9 , the relationship of Slope with Intercept is negative; that is, low Intercepts tend to correspond with steep Slopes. However, it is possible that individuals who obtained low Intercepts simply had more "room" to increase their reaction times within the 7-second time limit, thus increasing their Slope scores. Since high Intercept values were related to slower Grand Mean Reaction Times, as well as less accurate performance, and more "time-outs" occurred on the nine-digit items, it is likely that the 7-second time limit produced a ceiling effect.

The high positive correlation between the slope and accuracy suggests that performing accurately is related to increasing reaction time substantially as the stimuli increase in length. Steeper slopes also correspond with slower grand mean reaction times. These slower reaction times were also related to higher accuracy.

Table 5.9

Intercorrelations Among Perceptual Speed and Accuracy Test Scores

	Intercept	Slope	% Correct
Slope	$-.27^a$		
Percent Correct	$-.26^b$	$.64^a$	
Grand Mean[c]	$.35^a$	$.79^a$	$.45^a$

[a] $p \leq .001$

[b] $p \leq .003$

[c] Grand mean reaction time in this section refers to:

$$\text{Grand Mean} = \frac{\bar{X}_{2\text{-digits}} + \bar{X}_{5\text{-digits}} + \bar{X}_{7\text{-digits}} + \bar{X}_{9\text{-digits}} + \bar{X}_{\text{words}}}{5}$$

Correlations With Other Measures. The Perceptual Speed and Accuracy Test score that relates most highly with scores from the other computer-administered tests is the Intercept (see Table 5.3). Scores correlating most highly with the Intercept are the Choice Reaction Time Total and the Short-Term Memory Grand Mean Reaction Time.

The PS & A Grand Mean Reaction Time also correlates highly with scores from several of the computerized tests. Among the highest of these correlations are those with Target Identification Mean Reaction Time and the Short-Term Memory Grand Mean Reaction Time. The PS & A Slope correlated with accuracy on the Short-Term Memory Test but was not highly correlated with most of the other computer-administered measures.

The Perceptual Speed and Accuracy Intercept value correlates relatively highly with all of the cognitive paper-and-pencil measures (see Table 5.4). Its highest correlations were with Maze, which is a spatial scanning test ($r = -.57$), Orientation Test 1 ($r = -.5$,) and Reasoning Test 1 ($r = -.48$).

The Slope was most highly correlated with Reasoning Test 2 ($r = .27$). Accuracy on the PS & A test was most highly correlated with Reasoning Test 2 and Orientation Test 1 ($r = .31$), and Assembling Objects ($r = .30$). Object Rotation ($r = -.35$) and Maze ($r = -.33$) produced moderate correlations with the PS & A Grand Mean Reaction Time.

Generally speaking, the pattern of correlations for the Perceptual Speed and Accuracy scores is similar to that seen for the Memory Search Test. The PS & A Intercept and Grand Mean RT scores show patterns fairly similar to those for the same scores on the Memory Test, but PS & A Intercept shows a much stronger relationship with the cognitive, paper-and-pencil test scores than does the memory Intercept. Also, PS & A Slope generally shows lower correlations with all other measures as does the memory Slope.

Modifications for Fort Knox Field Test. Several changes were made to this test following the Fort Lewis pilot test. A reduction in the number of items was considered desirable in order to cut down the testing time, and the reliability of the test scores (see Table 5.8) indicated that the test length could be considerably reduced without causing the reliabilities to fall below acceptable levels. Item deletion was accomplished in two ways. First, all the seven-digit items were deleted (16 items). Examination of Figure 5.7 shows that such deletions should have little effect on the test scores, since the relationship between number of digits and reaction time is linear, and the items containing two, five, and nine digits should provide sufficient data points. Second, 16 more items were deleted by deleting four items from each of the remaining three digit categories (two, five, and nine) and from the "word" items. The following factors were considered in selecting items for deletion:

- Item intercorrelations within stimulus type and digit size were examined. In many cases, one item did not correlate highly with the others. Items that produced the lowest intercorrelations were deleted. Use of this criterion resulted in 13 item deletions.
- When item intercorrelations did not differ substantially, accuracy rates and variances were reviewed but did not indicate any clear candidates for deletion.
- When all the above were approximately equal, the decision to retain an item was based on its correct response (i.e., "same" or "different"). If retaining the item would have caused an imbalance between the responses, it was deleted. This was, in effect, a random selection.

Deletion of the 32 items left a 48 item test.

Several other changes were made, either to correct perceived shortcomings or to otherwise improve the test. The symbolic nine-digit items were modified to make them more difficult. As previously noted, these items had originally been developed in such a way that the symbols were in "chunks," thus making the items, in effect, much shorter than the intended nine digits; these groups were broken up. Five items were changed so that the correct response was "different" rather than "same" in order to balance type of correct response within digit level. Finally, the time allowed to make a response to an item was increased from 7 seconds to 9 seconds in order to give subjects sufficient time to respond, especially for the more difficult items.

The revised test, then, contained 48 items; 36 were divided into 12 Type (alpha, numeric, symbolic, mixed) by Number of Digits (two, five, nine) cells, and 12 were word items.

We also changed the presentation of the items so that they disappeared from the display screen as soon as the subject released the "home" button. This was intended to correct the problem of confounding decision time with movement time that was discussed above.

Target Identification Test

Test Description. The Target Identification Test is a measure of perceptual speed and accuracy. The objects perceived are meaningful figures, however, rather than a series of numbers, letters, or symbols as in the the preceding test.

In this test, each item shows a target object near the top of the screen and three labeled stimuli in a row near the bottom of the screen. Examples are shown in Figure 5.9. The subject is to identify which of the three stimuli represents the same object as the target and to press as quickly as possible the button (blue, yellow, or white) that corresponds to that object. A flow chart indicating the series of events in this test is presented in Figure 5.10.

Five parameters were varied in depicting objects for the test. The first was type of object. The objects shown on the screen are based on military vehicles and aircraft as shown on the standard set of flashcards used to train soldiers to recognize equipment presently being used by various nations. We sorted these cards into four basic types: tanks and other tracked vehicles, fixed-wing aircraft, helicopters, and "wheeled" vehicles. Then we prepared computerized drawings of representative objects in each type. These drawings were not intended to be completely accurate renditions but rather to depict the figures in a less complex drawing while retaining the basic distinguishing features. Twenty-two drawings of objects were prepared.

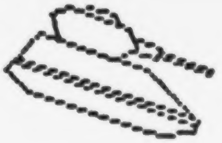
The second parameter was the position of the correct response--that is, on the left, middle, or right side of the screen. The third parameter was the orientation of the target object--whether it is "facing" in the same direction as the stimuli (the objects to be matched with the target) or in the opposite direction. This reduces to the target object "facing" left (one's left as one looks at the screen) or "facing" right.

The fourth parameter was the angle of rotation (from horizontal) of the target object. Seven different angular rotations were used for the Fort Lewis administration of this test: 0^o, 20^o, 25^o, 30^o, 35^o, 40^o, and 45^o. Example 1 in Figure 5.9 shows a rotated target object and Example 2 shows an unrotated object (0^o).

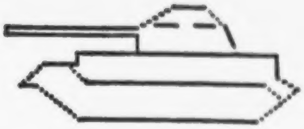
The fifth parameter was the size of the target object. Ten different levels of size reduction were used in the Fort Lewis administration: 40%, 50%, 55%, 60%, 65%, 75%, 80%, 85%, 90%, and 100%. Forty percent reduction means that the target object was 40 percent of the size of the stimulus objects at the bottom of the screen.

We had no intention of creating a test that had items tapping each cell of a crossed design for these five parameters. Instead, we viewed this tryout of the test as an opportunity to explore a number of different factors that could conceivably affect test performance. A total of 44

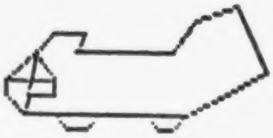
EXAMPLE 1.



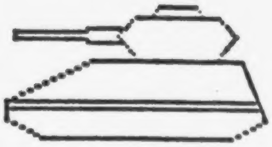
TARGET



BLUE

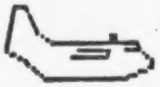


YELLOW

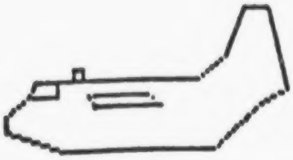


WHITE

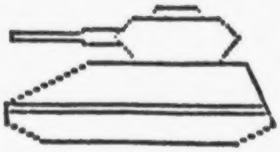
EXAMPLE 2.



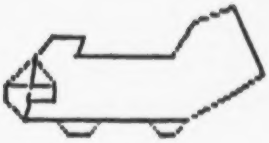
TARGET



BLUE



YELLOW



WHITE

Figure 5.9. Graphic displays of example items from the computer-administered Target Identification Test.

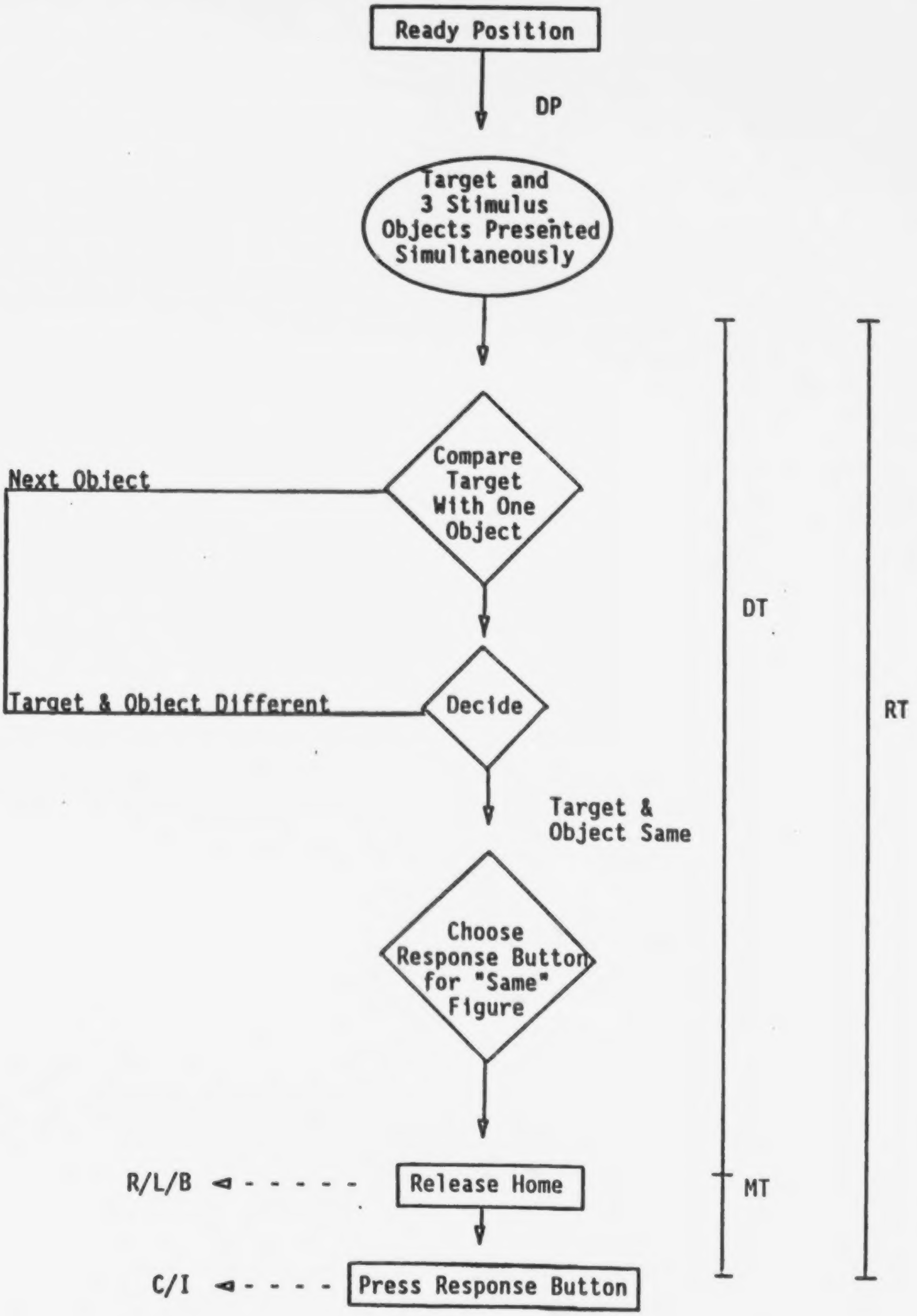

Ready Position
DP
Target and 3 Stimulus Objects Presented Simultaneously
Compare Target With One Object
Next Object
Decide
Target & Object Different
Target & Object Same
Choose Response Button for "Same" Figure
Release Home
R/L/B
Press Response Button
C/I
DT
MT
RT

Figure 5.10. Target Identification Test.

items were included on the test.

Test Characteristics. Table 5.10 shows data from the Fort Lewis pilot test of the Target Identification Test. With reference to the first part of the table, we see that the average time to read the instructions was about 2 minutes, with a range of 1.1 to 9.2 minutes. The time required to take the actual test averaged 3.6 minutes, and ranged from 3 to 5.5. Hence, the total test time (instruction plus actual test) ranged from 4.1 to 12.8 minutes and averaged 5.6.

The subject has 9.99 seconds to make a response on this test. Very few time-outs occurred, much less than one per person on the average, and with a maximum of two. The number of invalid responses was fairly high for this test, 3.2 on the average.

Dependent Variables. The primary dependent variables or scores for this test were Total Reaction Time (includes both decision and movement times) for correct responses, and the percent of responses that were correct. Total Reaction Time was used rather than decision time because it seems to be more ecologically valid (i.e., the Army is interested in how quickly a soldier can perceive, decide, and take some action and not just in the decision time). Also, various analyses of variance, discussed below, showed similar results for the two measures.

The second part of Table 5.10 shows data from the two dependent measures of concern: Total Reaction Time and Percent Correct. The test was conceived as a speeded test, in the sense that each item could be answered correctly if the subject took sufficient time to study the items and, therefore, the reaction time measure was intended to show the most variance. The data show that these intentions were achieved, since the mean Percent Correct was 92.6 with a standard deviation of 8.3, while the Reaction Time mean was 218 hundredths of a second with a standard deviation of 68.8. The reliability estimates show that the Reaction Time measure was highly reliable (.97), and it was about 20 points higher than the reliability for Percent Correct.

We performed a number of analyses of variance in order to investigate the effects of the five parameters described above on the most important dependent variable, Mean Total Reaction Time. Because of the number of parameters and levels within each parameter, a completely crossed design was not feasible. Instead, we carried out several one-way and two-way ANOVAs. Basically, the analyses showed that all the parameters had significant effects (well beyond the .01 level) on the mean reaction time score, but that many parameters included too many levels in the sense that there was little difference between scores for adjacent levels of a parameter. The results of these analyses were used to guide the revision of this test, described below.

Correlations With Other Measures. Correlations between Mean Reaction Time and Percent Correct on the Target Identification Test and scores on other Pilot Trial Battery tests were computed. Correlations of Mean Reaction Time with other computer tests ranged from .06 to .58 (see Table 5.3). The strongest relationships were with Perceptual Speed and Accuracy and Short-Term Memory, while the weakest were with several Simple and Choice Reaction Time measures. Percent correct correlated most highly with Short-

Table 5.10

Pilot Test Results from Fort Lewis: Target Identification Test (N = 112)

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	2.01	1.04	1.10 - 9.21
Time to Complete Test (minutes)	3.61	0.45	2.96 - 5.46
Total Test Time (minutes)	5.62	1.23	4.12 - 12.81
Time-Outs (per person)	.06	.28	0 - 2
Invalid Responses (per person)	3.20	3.62	0 - 29

Dependent Measures	Mean	SD	Range	R_{xx}[a]
Total Reaction Time[b]	218.51	68.75	113.10 - 492.95	.97
Percent Correct	92.60	8.30	34.1 - 100	.78

[a] Reliability estimates computed using odd-even procedure with Spearman-Brown correction.

[b] In hundredths of a second.

Term Memory (r = .51 with Percent Correct) and Perceptual Speed and Accuracy Slope (r = .27). The lowest relationships were with the reaction time measures and two measures on the Short-Term Memory Test (r = .07 with Slope and .05 with Grand Mean).

For Mean Reaction Time, correlations ranged from -.30 to -.50 with paper-and-pencil tests (see Table 5.4). The strongest relationships were with the Maze Test and Orientation Test 1; the weakest were with Assembling Objects and Path.

Percent Correct correlations with paper-and-pencil tests ranged from .11 to .29, the lowest being with Orientation Tests 1 and 3, and the highest with Assembing Objects and Maze.

Modifications for Fort Knox Field Test. Two parameters of the test were left unchanged: position of the correct response or object that "matched" the target (left, middle, or right position) and direction in which the target object faced (in the same or opposite direction of the objects to be compared). Analyses of the Fort Lewis data indicated that opposite-facing targets appeared to be a bit more difficult (i.e., had higher mean reaction times), and data on object position showed that those in the middle were slightly "easier" (faster reaction time). We thought it best, however, to balance the items with respect to these two parameters in order to control the response style.

The other three parameters were changed. The objects to be matched with the target were made to be all from one type (helicopters or aircraft or tanks, etc.) or from two types, rather than from one, two, or three. This was done because analyses showed the "three-type" items to be extremely easy. Rotation angles were reduced from seven levels to just two, 0° and 45°, since analyses showed that angular rotations near 0° had very little effect on reaction time.

Finally, the size parameter was radically changed. The target object was either 50 percent of the stimulus objects, or was made to "move." The "moving" items were made to initially appear on the screen as a small dot, indistiguishable, and to then quickly and successively disappear and reappear, slightly enlarged in size and slightly to the left or right (depending on the side of the screen on which the target initially appeared) of the prior appearance. Thus, the subject had to observe the moving and growing target until certain of matching it to one of the stimulus objects. These "moving" items were thought to represent greater ecological or content validity, but still to be a part of the perception construct.

The revised test consisted of 48 items, distributed one each in the 48 cells depicted in Figure 5.11.

	Left-Facing								Right-Facing							
	One Type				Two Types				One Type				Two Types			
	50%		Moving		50%		Moving		50%		Moving		50%		Moving	
	0°	45°	0°	45°	0°	45°	0°	45°	0°	45°	0°	45°	0°	45°	0°	45°
Correct: Left Object																
Correct: Middle Object																
Correct: Right Object																

Figure 5.11. Distribution of 48 items on the revised Target Identification Test according to five parameters.

PSYCHOMOTOR PRECISION

This construct is the ability to make muscular movements necessary to adjust or position a machine control mechanism. This ability applies to both anticipatory movements (i.e., where the subject must respond to a stimulus condition which is continuously changing in an unpredictable manner) and controlled movements (i.e., where the subject must respond to a stimulus condition which is changing in a predictable fashion, or making only a relatively few discrete, unpredictable changes). Psychomotor precision thus encompasses two of the ability constructs identified by Fleishman and his associates, control precision and rate control (Fleishman, 1967).

Performance on tracking tasks is very likely related to psychomotor precision. Since tracking tasks are an important part of many Army MOS, development of psychomotor precision tests was given a high priority. The Fort Lewis computer-administered battery included two measures for pilot testing this ability.

Target Tracking Test 1

The Target Tracking Test 1 was designed to measure subjects' ability to make fine, highly controlled movements to adjust a machine control mechanism in response to a stimulus whose speed and direction of movement are perfectly predictable. Fleishman labeled this ability control precision.

During World War II, Army Air Force researchers working in the Aviation Psychology Program used several control precision tests in an attempt to predict performance for several aircrew jobs (Melton, 1947). The test which proved to be the most valid predictor was the Rotary Pursuit Test. In this test the subject is presented with a round metal target which revolves near the edge of a phonograph-like disk. The subject is given a metal stylus and told to maintain contact with the target as it rotates. The Rotary Pursuit Test served as a model for Target Tracking Test 1.

Test Description. For each trial of this pursuit tracking test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box, and centered in the box is a crosshair. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject's task is to keep the crosshairs centered within the target at all times. The subject uses a joy stick, controlled with one hand, to control movement of the crosshairs. Figure 5.12 presents a schematic representation of this task.

Several item parameters were varied from trial to trial. These include the speed of the crosshairs, the maximum speed of the target, the difference between crosshairs and target speeds, the total length of the path, the number of line segments comprising the path, and the average amount of time the target spends traveling along each segment. Obviously, these parameters are not all independent; for example, crosshairs speed and maximum target speed determine the difference between crosshairs and target speeds.

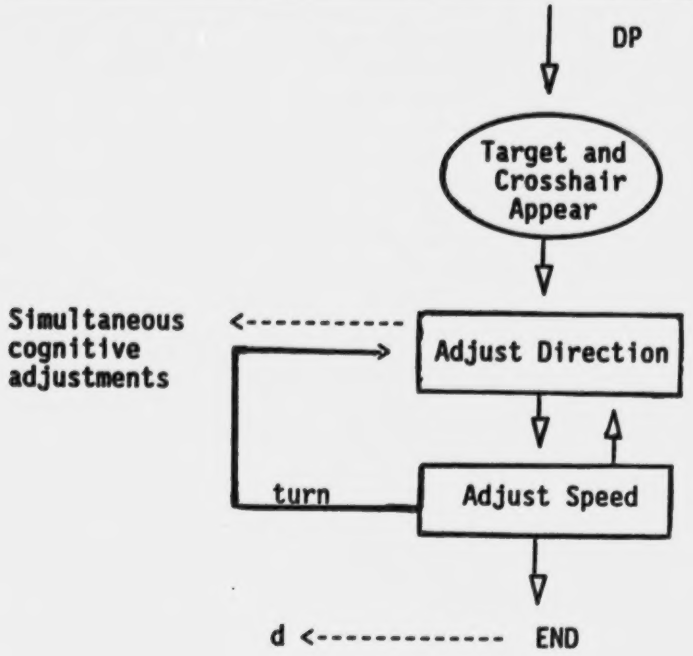

DP
Target and Crosshair Appear
Simultaneous cognitive adjustments
Adjust Direction
turn
Adjust Speed
d
END

Figure 5.12. Target Tracking Test 1

For the Fort Lewis battery, subjects were given 18 test trials. Three of the 18 paths were duplicates (the paths for trials 15-17 were identical to the paths for trials 1, 2, and 7). Except for these duplicates, the test was constructed so that the trials at the beginning of the test were easier than trials at the end of the test. In other words, target and crosshairs speeds were slower during the first several trials than during the final trials, the paths were shorter, the paths included fewer line segments, and so forth.

Dependent Measures. Two classes of dependent measures were investigated for this test: (1) tracking accuracy, and (2) improvement in tracking performance, based on the three duplicate paths included in the test.

Two tracking accuracy measures were investigated, time on target and distance from the center of the crosshairs to the center of the target. Kelley (1969) demonstrated that distance is a more reliable measure of tracking performance than time on target. Therefore, the test program computes the distance[4] from the crosshairs to the center of the target several times each second, and then averages these distances to derive an overall accuracy score for that trial. Subsequently, when the distribution of subjects' scores on each trial was examined, it was found that the distribution was highly positively skewed. Consequently, the trial score was transformed by taking the square root of the average distance. As a result, the distribution of subjects' scores on each trial was more nearly normal. These trial scores were then averaged to determine an overall tracking accuracy score for each subject.

---

[4] The COMPAQ video screen is divided into 200 pixels vertically and 640 pixels horizontally, with each vertical pixel equivalent to three horizontal pixels. All distance measures were computed in horizontal pixel units.

Prior to the Fort Lewis pilot test, it was expected that subjects' tracking proficiency would improve considerably over the course of the test. That was one of the reasons that initial test trials were designed to be easier than the later test trials. However, analyses of the Fort Lewis data revealed that subjects' performance on trials 1, 2, and 7 actually differed little from their performance on the duplicate trials 15-17. Therefore, it was decided that no further measure of improvement in tracking performance would be computed.

Test Characteristics. Table 5.11 presents data for Target Tracking Test 1 based on the Fort Lewis pilot test. The 18 trials of the test required 9 minutes to complete. Since all subjects received the same set of paths, there was virtually no variability. Instruction time mean was 1.2 minutes. The range of total test time was from 9.4 to 12.2 minutes, with a mean of 10.3 minutes.

Mean and standard deviation for overall accuracy score were 1.44 and .45, respectively. As a result of the square root transformation, the distribution of accuracy scores was only slightly positively skewed. The internal consistency reliability of the accuracy score, computed by comparing the mean accuracy scores for odd and even trials, was .97.

Table 5.11

Pilot Test Results From Fort Lewis: Target Tracking Test 1 (N - 112)

Descriptive Characteristics	Mean	SD	Range	
Time to Read Instructions (minutes)	1.20	.43	.33 - 3.09	
Time to Complete Test (minutes)	9.07	.02	9.05 - 9.12	
Total Test Time (minutes)	10.27	.43	9.42 - 12.17	
Dependent Measure	**Mean**	**SD**	**Range**	r_{xx}[a]
Distance[b]	1.44	.45	.95 - 3.40	.97

[a] Spearman-Brown corrected split-half reliability for odd-even trials.

[b] Square root of the average within-trial distance (horizontal pixels) from the center of the target to the center of the cross-hairs, averaged across all 18 trials (or items) on the test.

Four one-way analyses of variance were executed to determine the effects on tracking accuracy of average segment length, average time required for the target to travel a segment, maximum crosshairs speed, and difference between maximum crosshairs speed and target speed. All four item parameters were significantly related to accuracy score, with crosshairs speed accounting for the most variance, and difference between target and crosshairs speed accounting for the least variance. It should be noted that all four parameters were highly intercorrelated (the six intercorrelations ranged from .37 to .87, with a median intercorrelation of .52), and all four were also correlated with trial number (i.e., trials were designed to become more difficult as the test progressed). As a result, it is difficult to interpret the results of these ANOVAs.

Correlations With Other Measures. Table 5.3 shows the correlations between the Target Tracking Test 1 and other computer-administered measures. The test was highly correlated with Target Tracking Test 2 ($r = .76$). Because that test was intended to be a measure of a different construct, multilimb coordination, this correlation is troubling. In part, it reflects the great similarity of these two tests; both used the same set of 18 tracking paths, presented in the same order. The only difference was in the type of control adjustments required; for Target Tracking Test 1 subjects used a joy stick operated with their preferred hand to make all control adjustments, and for Target Tracking Test 2 subjects used both hands to manipulate horizontal and vertical sliding resistors. It is probable that the large correlation is due mainly to the high degree of task similarity.

Target Tracking Test 1 was also significantly correlated with tracking performance on the other psychomotor test, the Target Shoot Test ($r = .32$ for Distance from the center of the crosshairs to the center of the target at the time of firing, $r = .43$ with percent of hits). The significant intercorrelations among the psychomotor tests reflect a general psychomotor ability factor. (This factor also emerged in a factor analysis of the computer tests, discussed below.)

Correlations with Target Tracking Test 1 also exceeded .30 for four other computer-dependent measures--Perceptual Speed and Accuracy Intercept ($r = .36$), Target Identification Mean Reaction Time ($r = .46$), and Total Reaction Time for the Simple and Choice reaction time tests ($r = .31$ and .39, respectively). These measures all reflect the speed of rather basic cognitive processes (e.g., detection, comparison).

Target Tracking Test 1 also correlated significantly with all the cognitive paper-and-pencil tests in the pilot trial battery (Table 5.4). These correlations ranged from .27 with the Assembling Objects Test to .52 with the Maze Test. As noted previously, most of these paper-and-pencil tests were designed to measure some aspect of spatial ability. In the literature review for the psychomotor ability domain, it was shown that control precision correlated more highly with spatial ability than with any other cognitive ability. Thus, the significant correlations between Target Tracking Test 1 and the paper-and-pencil tests do not represent a surprise.

Modifications for Fort Knox Field Test. Several changes were made in the paths comprising this test for the Fort Knox field test. First, all paths were modified so that each would run for the same amount of time (approximately .36 minute). The primary reason for this change was that the program computes distance between the crosshairs and target a set number of times each second. If all paths run the same amount of time, then the accuracy measure for all trials will be based on the same number of distance assessments.

Second, three item parameters were identified to direct the format of test trials: maximum crosshairs speed, difference between maximum crosshairs speed and target speed, and number of path segments. Given these parameters and the constraint that all trials run a fixed amount of time, the values of all other item parameters (e.g., target speed, total length of the path) can be determined. Three levels were identified for each of the three parameters. These were completely crossed to create a 27-item test, and items were then randomly ordered. These procedures for item development should alleviate pilot testing problems in interpreting test results in light of correlated item parameters.

Third, in spite of these changes, which added 50 percent more trials to the test, testing time was actually reduced slightly (25 seconds less, it was estimated) because of the standardization of the trial time.

Target Shoot Test

The Target Shoot Test was modeled after several compensatory and pursuit tracking tests used by the AAF in the Aviation Psychology Program (e.g., the Rate Control Test). The distinguishing feature of these tests is that the target stimulus moves in a continuously changing and unpredictable speed and direction. Thus, the subject must attempt to anticipate these changes and respond accordingly.

Test Description. For the Target Shoot Test, a target box and crosshairs appear in different locations on the computer screen. The target moves about the screen in an unpredictable manner, frequently changing speed and direction. The subject controls movement of the crosshairs via a joy stick. The subject's task is to move the crosshairs into the center of the target. When this has been accomplished, the subject must press a button on the response pedestal to "fire" at the target. The subject's score on a trial is the distance from the center of the crosshairs to the center of the target at the time the subject fires. The test consists of 40 trials. A schematic depiction of these trials is presented in Figure 5.13.

Several item parameters were varied from trial to trial. These parameters included the maximum speed of the crosshairs, the average speed of the target, the difference between crosshairs and target speeds, the number of changes in target speed (if any), the number of line segments comprising the path of each target, and the average amount of time required for the target to travel each segment. These parameters are not all independent, of course. Moreover, the nature of the test creates a problem in characterizing some trials; a trial terminates as soon as the subject fires at the target, so one subject may see only a fraction of the line segments, target speeds, etc., that another subject sees.

Dependent Variables. Three dependent measures were obtained for each trial. Two were measures of firing accuracy: (1) the distance from the

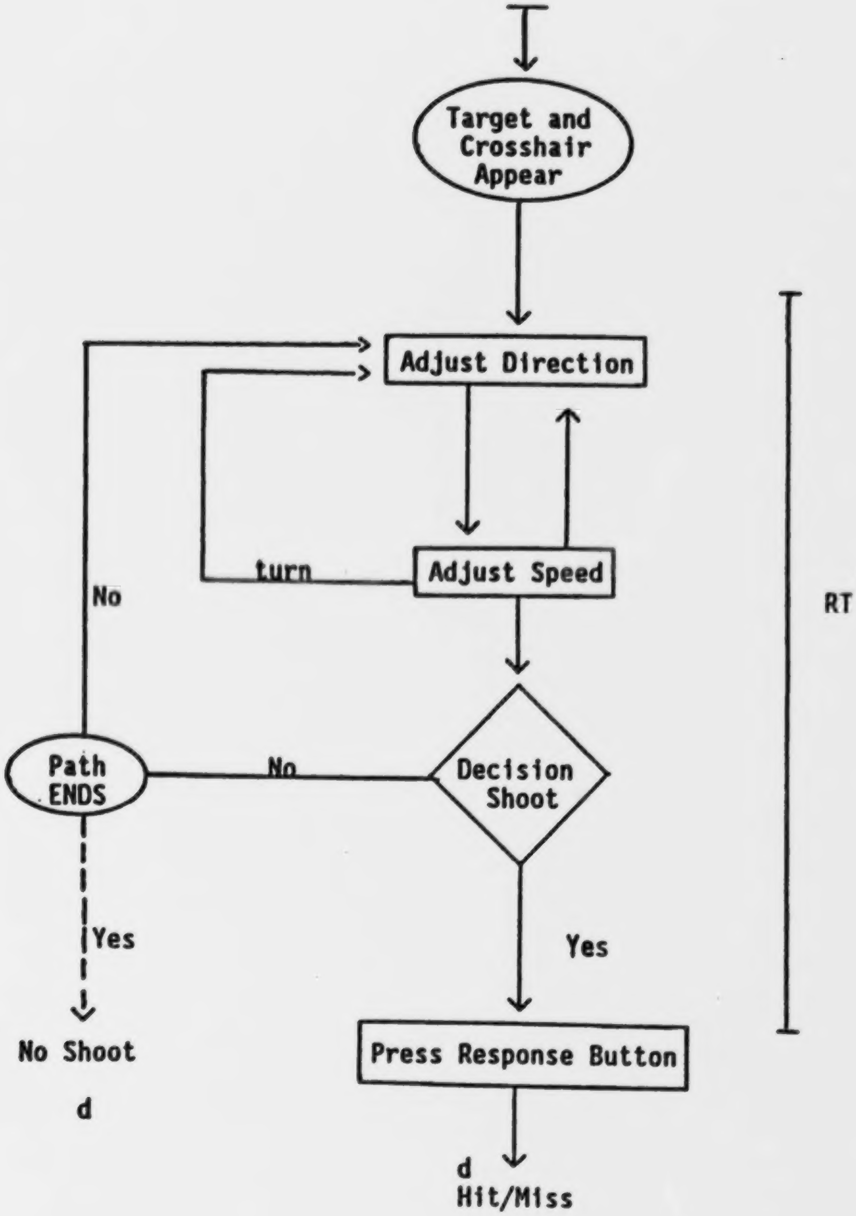

Target and Crosshair Appear
Adjust Direction
Adjust Speed
turn
No
Path ENDS
No
Decision Shoot
Yes
Yes
No Shoot
d
Press Response Button
d
Hit/Miss
RT

Figure 5.13. Target Shoot Test.

center of the crosshairs to the center of the target at the time of firing, and (2) whether the subject "hit" or "missed" the target. The two were very highly correlated. However, the former provides quite a bit more information about firing accuracy than the latter, so Distance was retained as the accuracy measure. Distances were averaged across trials to obtain an overall accuracy score. In some trials, the subject failed to fire at the target so no distance score was obtained; those trials were not included in the overall test accuracy score.

The third dependent measure was a speed measure, representing the time from trial onset until the subject fired at the target. Again, trials were omitted if the subject failed to fire a shot. This last measure was not used in any subsequent analyses, primarily because we had no clear idea about how to view its relationship to the construct being measured on this test, or to constructs measured on other tests.

Test Characteristics. Table 5.12 presents data based on the Fort Lewis pilot test. The total time for this test averaged close to 4 minutes, with about 1.6 minutes for instructions and 2.2 minutes for the test itself. In two or three trials, on the average, a subject failed to fire at the target.

Split-half reliability across odd-even trials was .93 for Mean Distance and .78 for Percent Hits. The average Percent of Hits was 58, with a range from 0 to 83. These results show that the Distance score is highly reliable and has adequate variance, and the Percent of Hits score is acceptably reliable and also has adequate variance. Also, the 58 percent mean on this score shows that the test was at about the right level of difficulty.

Analyses of variance were executed to determine the effects of several item parameters (crosshairs speed, average target speed, and average segment length) on mean distance. All were found to be related to item difficulty. However, interpretation of these results was made difficult by the correlations among the parameters and by item order effects (i.e., the last dozen or so trials presented the most difficult tracking problems).

Correlations With Other Measures. Correlations with other computer-administered tests exceeded .30 only for the two tracking tests (Table 5.3). The correlation was actually higher with Tracking Test 2 ($r = .47$ versus .32 for Tracking Test 1), possibly indicating that performance on the Target Shoot Test is influenced by multilimb coordination.The Target Shoot Test Mean Distance was relatively uncorrelated with cognitive paper-and-pencil test scores (Table 5.4). The highest correlation was -.23, with the Maze Test. Thus, it was felt that the test was not heavily dependent upon any spatial-perceptual abilities.

Modifications for Fort Knox Field Test. Because of its high reliability and its independence from other ability measures, the test was not modified for Fort Knox field testing.

Table 5.12

Pilot Test Results From Fort Lewis: Target Shoot Test (N = 112)

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	1.58	.61	.51 - 5.10
Time to Complete Test (minutes)	2.22	.23	1.81 - 3.29
Total Test Time (minutes)	3.80	.68	2.71 - 7.58
No. of Trials Without Firing[a]	2.77	3.97	0 - 40

Dependent Measures	Mean	SD	Range	r_{xx}[b]
Distance[c]	2.83	.52	1.93 - 7.03	.93
Percent of hits[a]	58	13	0 - 83	.78

---

[a] One subject failed to fire at any targets. Excluding this subject, mean, SD, and range for number of trials without firing were 2.43, 1.78, and 0-8, respectively; mean, SD, and range for percent of hits were 59, 12, and 13-83, respectively.

[b] Spearman-Brown corrected split-half reliability for odd-even trials.

[c] Square root of the distance (horizontal pixels) from the center of the target to the center of the crosshairs at the time of firing, averaged across all trials in which the subject fired at the target. (There were a total of 40 trials or times on the test.)

MULTILIMB COORDINATION

The multilimb coordination construct reflects the ability to coordinate the simultaneous movement of two or more limbs. This ability is general to tasks requiring coordination of any two limbs (e.g., two hands, two feet, one hand and one foot). The ability does not apply to tasks in which trunk movement must be integrated with limb movements. It is most common in tasks where the body is at rest (e.g., seated or standing) while two or more limbs are in motion.

In the past, measures of multilimb coordination have shown quite high validity for predicting job and training performance, especially for pilots (Melton, 1947).

Target Tracking Test 2

Target Tracking Test 2 is modeled after a test of multilimb coordination developed by the AAF, the Two-Hand Coordination Test. This test required subjects to perform a pursuit tracking task in which horizontal and vertical movements of the target-follower were controlled by two handles. Validities of this test for predicting AAF pilot training success were mostly in the .30s (Melton, 1947).

Test Description. Target Tracking Test 2 is very similar to the Two-Hand Coordination Test. For each trial of Target Tracking Test 2, subjects are shown a path consisting entirely of vertical and horizontal lines. At the beginning of the path is a target box, and centered in the target box is a crosshairs. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject manipulates two sliding resistors to control movement of the crosshairs; one resistor controls movement in the horizontal plane, and the other in the vertical plane. The subject's task is to keep the crosshairs centered within the target at all times. Figure 5.14 contains a schematic depiction of the test.

This test and Target Tracking Test 1 are identical except for the nature of the required control manipulations. For Target Tracking Test 1 crosshairs movement is controlled via a joy stick, while for Target Tracking Test 2 crosshairs movement is controlled via the two sliding resistors. For the Fort Lewis battery, the same 18 paths were used in both tests, and the value of the crosshairs and target speed parameters was the same. The only other difference between the two tests was that subjects were permitted three practice trials for Target Tracking Test 2.

Dependent Variable. The same dependent measure or score was used for this test as for Tracking Test 1 (i.e, the square root of the average within-trial distance from the center of the crosshairs to the center of the target, averaged across all trials).

Test Characteristics. The 18 trials of the test (Table 5.13) required 9 minutes to complete. Since all subjects received the same set of paths, there was virtually no variability. Instruction time mean was 3.6. The range of total test time was from 11.5 to 15.5 minutes, with a mean of 12.7 minutes.

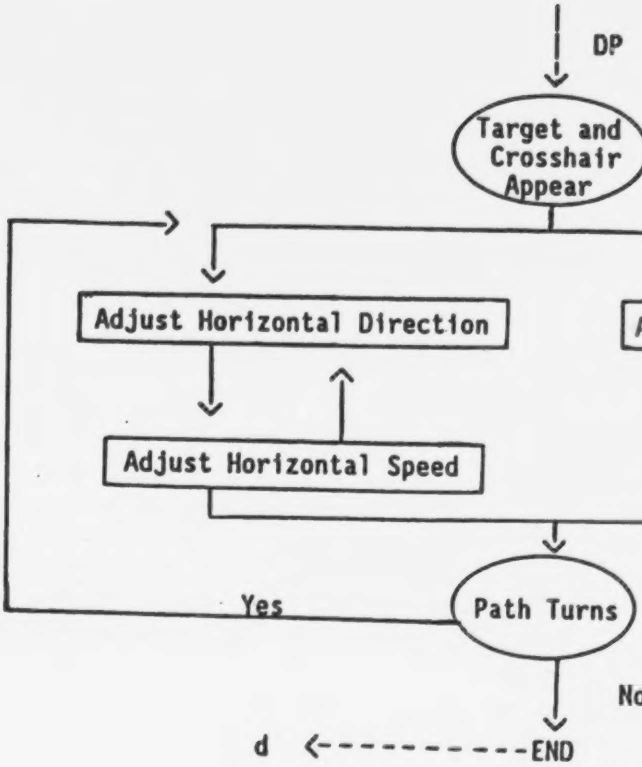

DP
Target and Crosshair Appear
Adjust Horizontal Direction
Adjust Vertical Direction
Adjust Horizontal Speed
Adjust Vertical Speed
Path Turns
Yes
No
d
END

Figure 5.14. Target Tracking Test 2.

Table 5.13

Pilot Test Results From Fort Lewis: Target Tracking Test 2 (N = 112)

Descriptive Characteristics	Mean	SD	Range	
Time to read instructions[a]	3.58	.68	2.39 - 6.38	
Time to complete test[a]	9.09	.02	9.03 - 9.13	
Total test time[a]	12.67	.68	11.50 - 15.48	
Dependent Measures	Mean	SD	Range	r_{xx}[a]
Distance[b]	2.02	.64	0 - 4.01	.97

---

[a] Spearman-Brown corrected split-half reliability for odd-even trials.

[b] Square root of the average within-trial distance (horizontal pixels) from the center of the target to the center of the crosshairs, averaged across all 18 trials (or items) on the test.

Mean and standard deviation for overall accuracy score were 2.02 and .64, respectively. As a result of the square root transformation, the distribution of accuracy scores was only slightly positively skewed. The internal consistency reliability of the accuracy score was .97. These results indicate that Target Tracking Test 2 is highly reliable as is Target Tracking Test 1, and that it is more difficult than is Target Tracking Test 1 (mean Distance score for Target Tracking Test 2 = 2.02 versus 1.44 for Target Tracking Test 1--a difference of about one standard deviation).

Four one-way analyses of variance were executed to determine the effects on tracking accuracy of average segment length, average time required for the target to travel a segment, maximum crosshairs speed, and difference between maximum crosshairs speed and target speed. All four item parameters were significantly related to accuracy score, with crosshairs speed accounting for the most variance and average segment length for the least. It should be noted again that all four parameters were highly intercorrelated (the six intercorrelations ranged from .37 to .87, with a median intercorrelation of .52), and all four were also correlated with trial number (i.e., items became more difficult as the test progressed). As a result, interpreting the results of these ANOVAs is difficult.

Correlations With Other Measures. Table 5.3 shows the correlations between the Target Tracking Test 2 and other computer-administered measures. The test was highly correlated with Target Tracking Test 1 (*r* = .76). Possible reasons for this correlation were discussed above (see Target Tracking Test 1).

Given the high correlation with Target Tracking Test 1, it would be expected that Target Tracking Test 2 would show a similar pattern of correlations with other computerized and paper-and-pencil ability measures. As Tables 5.3 and 5.4 show, this is essentially true. The only major difference is that Target Tracking Test 2 failed to correlate significantly with mean Total Response Time from the Simple Reaction Time Test (*r* = .11 versus *r* = .31 for Target Tracking Test 1).

Modifications for Fort Knox Field Test. Changes in Target Tracking Test 2 for the Fort Knox mirrored those made for Target Tracking Test 1. Test trials were changed completely. Test development was directed by three item parameters--number of segments, crosshairs speed, and difference between target and crosshairs speeds. The revised test includes 27 items. However, the items are no longer the same as those presented for Target Tracking Test 1, which should reduce the correlation between these tests.

NUMBER OPERATIONS

This construct involves the ability to perform, quickly and accurately, simple arithmetic operations such as addition, subtraction, multiplication, and division.

The current ASVAB includes a numerical operations test, containing 50 very simple arithmetic problems with a 3-minute time limit. Because of low item difficulty and the speeded nature of the test, correlations with other ASVAB subtests indicate that Numerical Operations is most strongly related to Coding --a measure of perceptual speed and accuracy. The present military-wide selection and classification battery, then, measures very basic number operations abilities which appear very similar to perceptual speed and accuracy abilities.

In the expert judgment process described in Chapter 1, this construct received a mean estimated validity of .40 with the highest value .44. The experts judged that this construct is an effective predictor of success in technical and clerical MOS. The authors, the scientific advisors, and the ARI scientists also thought that a computerized measure of this construct might prove superior to the paper-and-pencil format currently used.

The test designed to assess number operations abilities was not completed prior to the Fort Lewis pilot test, so no data are yet available to evaluate this measure. It has been prepared for administration as part of the test battery for the Fort Knox field test.

Number Memory Test

Test Description. This test was modeled after a number memory test developed by Dr. Raymond Christal at Air Force Human Resources Laboratory. The basic difference between the AFHRL test and the Number Memory Test concerns pacing of the number items. The former uses machine-paced presentation, while the latter involves self-paced item presentation. Both, however, require subjects to perform simple number operations such as addition, subtraction, multiplication, and division and both involve a memory task.

In the Number Memory Test, subjects are presented with a single number on the computer screen. After studying the number, the subject is to push a button to receive the next part of the problem. When the subject presses the button, the first part of the problem disappears and another number along with an operation term, such as Add 9 or Subtract 6, then appears. Once the subject has combined the first number with the second, he/she must press a button to receive the third part of the problem. Again, the second part of the problem disappears when the subject presses the button. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is true or false.

An example number operation item follows:

Item Set	8	
	+6	
	-3	
	x2	
	-4	
Probe	Is 16 the correct answer?	
Response	T White	F Blue

Figure 5.15 presents a flow chart for this test.

Test items vary with respect to number of parts--four, six, or eight--contained in the single item. Items also vary according to the delay between item part presentation or interstimulus delay period. One-half of the items include a brief delay (.5 second) while the other half contain a lengthier delay (2.5 seconds). The test contained 27 items.

This test is not a "pure" measure of number operations, since it also is designed to bring short-term memory into play. We decided that this was the most efficient way to proceed, since a second measure of short-term memory was thought desirable, at least at this point in the project.

Dependent Measures. Analyses planned for data that will be obtained from the Fort Knox field test administration include an investigation of the impact of item length (four, six, or eight) and interstimulus delay (.5 second or 2.5 seconds) on reaction time and percent correct, as well as comparisons of mean reaction time scores for item parts requiring addition, subtraction, multiplication and division. These analyses will be used to identify the dependent measures for scoring subject responses in the field test.

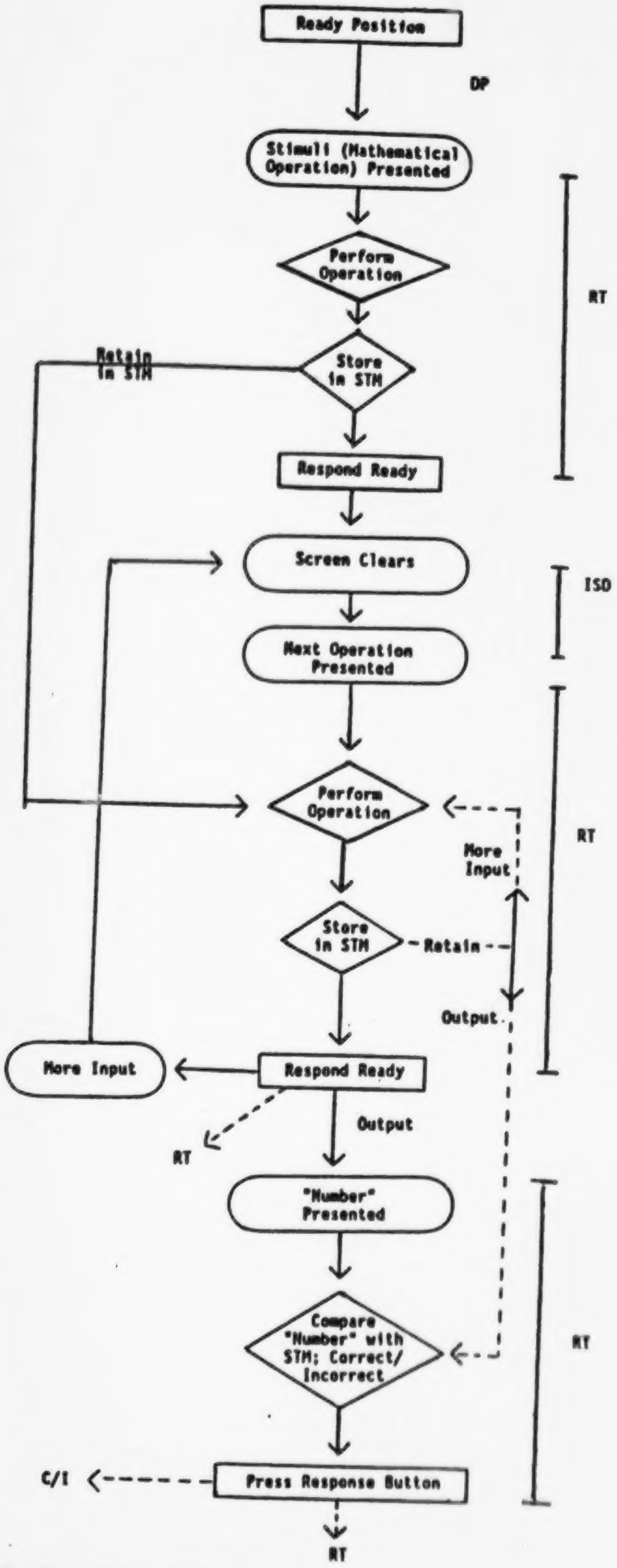

Ready Position
DP
Stimuli (Mathematical Operation) Presented
Perform Operation
RT
Retain in STM
Store in STM
Respond Ready
Screen Clears
ISD
Next Operation Presented
Perform Operation
More Input
RT
Store in STM
Retain
Output
More Input
Respond Ready
Output
RT
"Number" Presented
Compare "Number" with STM; Correct/Incorrect
RT
C/I
Press Response Button
RT

Figure 5.15. Number Memory Test

MOVEMENT JUDGMENT

Movement judgment is the ability to judge the relative speed and direction of one or more moving objects in order to determine where those objects will be at a given point in time and/or when those objects might intersect.

Movement judgment was not one of the constructs identified and targeted for test development by the literature review or expert judgments described in Chapter 1. However, a suggestion by Lloyd Humphreys, one of our scientific advisors, and the job observations we conducted at Fort Stewart, Ft. Bragg, Ft. Bliss, Ft. Sill, and Ft. Knox, led us to conclude that movement judgment was likely to be related to job performance in a number of combat MOS (e.g., 16S, 11B, 19D). Therefore, we decided to develop a movement judgment measure to be included in the Fort Knox field test.

Cannon Shoot Test

The Cannon Shoot Test measures subjects' ability to fire at a moving target in such a way that the shell that is fired hits the target when the target crosses the cannon's line of fire.

As part of its Aviation Psychology Program, the Army Air Force became interested in motion, distance, and orientation judgment and instituted development of a battery of motion picture and photograph tests (Gibson, 1947). One of the AAF measures was called the Estimate of Relative Velocities Test, a paper-and-pencil test. Each trial consisted of four frames. In each frame, two objects (airplanes) were shown flying along the same path in the same direction. In each subsequent frame, the trailing plane edged nearer the lead plane. The subject's task was to indicate on the final frame where the planes would intersect. Validities of this test for predicting pilot training success averaged approximately .18 (Gibson, 1947).

The present test was designed to test the construct that seems to underly the Estimate of Relative Velocities Test.

Test Description. At the beginning of each trial, a stationary cannon appears on the video screen, with the position of this cannon varying from trial to trial. The cannon is "capable" of firing a shell, which travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a response button to fire the shell in such a way that the shell intersects the target when the target crosses the cannon's line of fire. Figure 5.16 shows a flow chart for this test.

Three parameters determine the nature of each test trial. The first is the angle of the target movement relative to the position of the cannon; 12 different angles were used. The second is the distance from the cannon to the impact point (i.e., the point at which the target crosses the cannon's line of fire); four different distance values were used. Finally, the third parameter is the distance from the impact point to the fire point

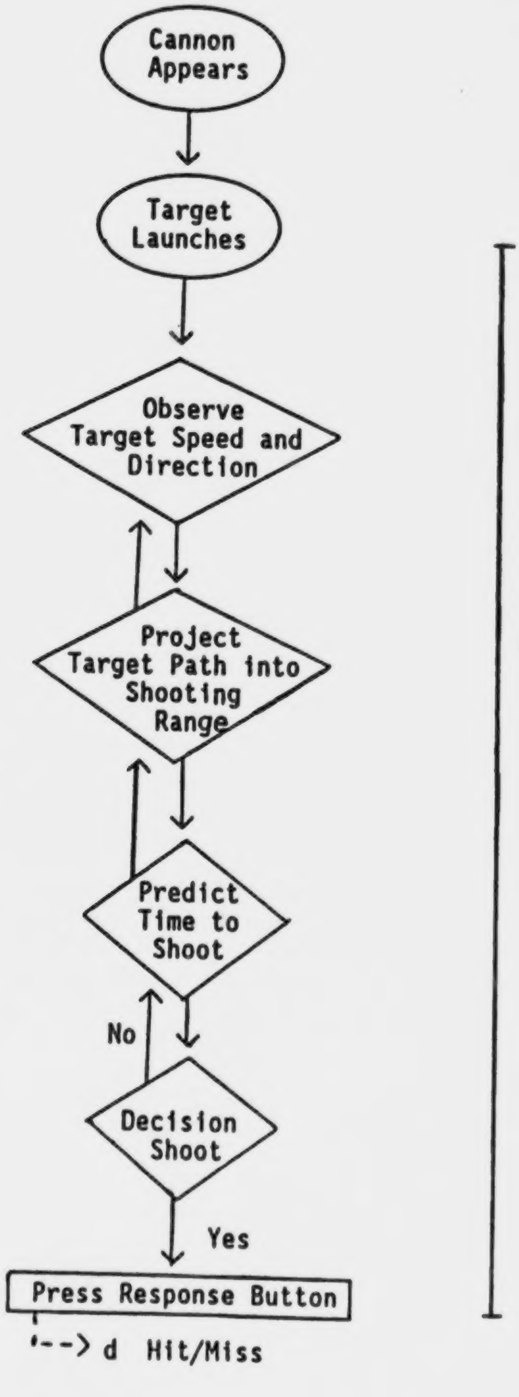

Cannon Appears
Target Launches
Observe Target Speed and Direction
Project Target Path into Shooting Range
Predict Time to Shoot
No
Decision Shoot
Yes
Press Response Button
--> d Hit/Miss
RT

Figure 5.16. Cannon Shoot Test.

(i.e., the point at which the subject must fire the shell in order to hit the center of the target); there were also four values for this distance parameter.

If a completely crossed design were used, it would necessitate a minimum of 192 trials (i.e., 12 x 4 x 4 = 192). Instead, a Latin square design was employed, so that the version of the test for the Fort Knox field test includes only 48 trials.

Dependent Measures. Three dependent measures are assessed on each trial. These include: (1) whether the shell hits or misses the target; (2) the distance from the shell to the center of the target at the time the target crosses the impact point; and (3) the distance from the center of the target to the fire point at the time the shell is fired. The Fort Knox data will be analyzed to determine which of these three measures is most reliable. Since the three will be highly intercorrelated, in the end it is likely that only one of the three will be retained as a dependent measure.

Test Characteristics. Prior to the Fort Knox Field Test, only minimal preliminary data are available for this test since it was not part of the Fort Lewis pilot test. It appears that the test will take approximately 12 minutes to complete, including instructions. It also appears that all three item parameters are related to item difficulty. That is, targets are more difficult to hit if the angle of the target is greater than 90% (i.e., the target is moving away from, rather than toward, the cannon), the impact point is far from the cannon, or the fire point is far from the impact point. Thus, targets that move rapidly are more difficult to hit than those that move slowly. However, all of these findings are based on observations of only a few subjects and are therefore tentative.

SUMMARY

Table 5.14 shows the means, standard deviations, and split-half reliabilities for 24 scores computed from the eight computer-administered tests which were pilot tested at Fort Lewis. As referred to throughout this chapter, Tables 5.3 and 5.4 show the intercorrelations between computer test scores, and the correlations between computer test scores and cognitive test scores. We make no further comment here since these data have been thoroughly discussed throughout the chapter.

Investigation of Machine Effects

One concern we had prior to the Fort Lewis pilot test was the extent to which computer measure scores would be affected by differences between testing stations. A testing station is one Compaq computer and the associated response pedestal; six such testing stations were used at Fort Lewis. As we mentioned in Chapter 1, differences across testing apparatus and unreliability of testing apparatus had been a problem in World War II psychomotor testing and thereafter. The recent advent of microprocessor technology was viewed as alleviating such problems, at least to some degree.

We ran some analyses of variance to provide an initial look at the extent of this problem with our testing stations. Thirteen one-way ANOVAs were run with testing stations as levels and computer test scores as the dependent variables. We ran separate ANOVAs for white males and non-white males in order to avoid confounding the results with possible subgroup differences. Also, only five testing stations were used since one station did not have enough subjects assigned to it. These results are shown in Table 5.15.

Of the 26 ANOVAs, only one reached significance at .05 level, about what would be expected by chance. These results were heartening. (Note that the distance measures in Table 5.15 have not been converted to the mean square root units; these are the sums of the mean distances across all items.).

One reason for these results was the use of calibration software. This software adjusted for the idiosyncratic differences of each response pedestal, insuring a more standardized test administration across testing stations.

Pilot Test Results: Comments

The results of the Fort Lewis pilot test of the computer-administered measures in the Pilot Trial Battery were extremely useful. The results showed very high promise for these measures in several ways:

1. The battery proved to be basically self-administering. The testing stations and battery software were successful in that almost every soldier could complete the entire battery with no assistance from the test monitor.
2. Only one testing station experienced equipment problems during the week of testing, showing that fairly large-scale testing

Table 5.14

Means, Standard Deviations, and Split-Half Reliability Coefficients for 24 Computer Measure Scores Based on Fort Lewis Pilot Test Data (N = 112)

	Mean	SD	Split-Half[a] Reliability
SIMPLE REACTION TIME (10 Items)			
Mean Decision Time (hs)[b]	29.25	8.10	.92
Mean Total Reaction Time (hs)	55.92	13.86	.94
Trimmed Standard Deviation (hs)	11.79	16.80	.66
Percent Correct	99	3	-.01
CHOICE REACTION TIME (15 Items)			
Mean Decision Time (hs)	36.78	7.75	.94
Mean Total Reaction Time (hs)	65.98	10.39	.91
Standard Deviation (hs)	8.92	3.75	.10
Percent Correct	99	3	-.16
DIFFERENCE IN SIMPLE & CHOICE REACTION TIME			
Decision Time (hs)	7.68	8.79	.86
Total Time (hs)	10.37	11.15	.79
SHORT-TERM MEMORY (50 Items)			
Intercept (hs)	97.53	30.28	.84
Slope (hs)	7.19	6.14	.54
Percent Correct	90	10	.95
Grand Mean (hs)	119.05	29.84	.88
PERCEPTUAL SPEED & ACCURACY (80 Items)			
Intercept (hs)	89.37	36.48	.85
Slope (hs)	33.14	9.78	.89
Percent Correct	87	8	.81
Grand Mean (hs)	294.22	57.13	.97
TARGET IDENTIFICATION (44 Items)			
Mean Total Time (hs)	218.51	68.75	.97
Percent Correct	93	8	.78
TARGET TRACKING 1 (18 Items)			
Mean Distance ($m\sqrt{m \text{ pixels}}$)[c]	1.44	.45	.97
TARGET TRACKING 2 (18 Items)			
Mean Distance ($m\sqrt{m \text{ pixels}}$)	2.01	.64	.97
TARGET SHOOT (40 Items)			
Mean Total Distance ($m\sqrt{m \text{ pixels}}$)	2.83	.52	.93
Percent "Hits"	58	13	.78

---

[a] Odd-even item correlation corrected to full test length with the Spearman-Brown formula.

[b] hs = hundredths of seconds.

[c] $m\sqrt{m \text{ pixels}}$ = mean of the square root of the mean distance from target, computed across all trials.

Table 5.15

Results of Analyses of Variance for Machine Effects: White and Non-White Males, Fort Lewis Sample

Test	White Males				Non-White Males			
	N	Mean	SD	F^a	N	Mean	SD	F^b
Reaction Time 1: Total RT (hsec)	45	58.29	30.17	0.79	26	58.58	12.44	0.22
Reaction Time 2: Percent Correct	45	98.91	2.57	1.13	26	97.84	3.29	1.14
Reaction Time 2: Total RT (hsec)	45	63.22	8.57	0.35	26	67.58	12.45	2.43
Memory: Percent Correct	45	90.89	5.75	0.46	26	85.54	12.82	0.51
Memory: Grand Mean (hsec)	45	110.13	22.45	0.16	26	118.00	30.38	1.49
Perceptual Speed & Accuracy: Percent Correct	45	85.84	5.85	0.75	26	79.50	9.91	0.29
Perceptual Speed & Accuracy: Grand Mean (hsec)	45	287.96	53.92	0.94	26	274.58	73.93	0.45
Identification: Percent Correct	45	94.00	4.60	0.21	26	90.54	9.46	0.87
Identification: Total RT (hsec)	45	190.02	49.24	1.13	26	208.62	57.67	1.59
Tracking 1: Distance	45	1548.31	458.60	1.41	26	2608.58	1567.33	0.42
Tracking 2: Distance	45	3410.29	1864.34	2.61*	26	5161.27	2740.69	1.42
Target Shoot: Percent Hits	45	63.22	9.35	0.37	25	58.88	10.75	0.10
Target Shoot: Distance	45	789.71	153.93	0.44	25	915.12	311.22	0.22

[a] Degrees of freedom = 4, 40; *F* for alpha = .05 is 2.60

[b] Degrees of freedom = 4, 21; *F* for alpha = .05 is 2.87

* Significant at = .05

with portable computer equipment is feasible.

3. The measures showed acceptable psychometric properties, although there was definitely room for improvement in several cases. The analyses were instructive for making these changes.

4. The soldiers liked the test battery. Virtually every soldier expressed a preference for the computer-administered tests compared to the paper-and-pencil tests. We thought there were several reasons for this attitude: novelty; the game-like nature of several tests; and the fact that the battery was, in large part, self-paced, allowing each soldier to thoroughly understand the instructions and to work through the battery at his/her own speed.

Chapter 5 References

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. *Human Factors, 9*, 1017-1032.

Gibson, J. J. (Ed.) (1947). Motion picture testing and research. *Army Air Forces Aviation Psychology Research Program Reports, 7*, Washington, D.C.: Government Printing Office.

Jensen, A. R. (1982). Reaction time and psychometric g. In M. J. Eysenck (Ed.), *A model for intelligence*, Springer-Verlag.

Kelley, Charles R. (1969). The Measurement of Tracking Proficiency. *Human Factors, 11*, 43-64.

Keyes, M. A. (1985, in press). *A review of the relationship between reaction time and mental ability*. Minneapolis, MN: Personnel Decisions Research Institute.

Melton, A. W. (Ed.) (1947). *Apparatus tests* (Army Air Forces Aviation Psychology Program Research Report No. 4). Washington, D.C.: U.S. Government Printing Office.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153*, 652-654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica, 30*, 276-315.

Thorson, G., Hochhaus, L., & Stanners, R. F. (1976). Temporal changes in visual and acoustic codes in a letter-matching task. *Perception and Psychophysics, 19*, 346-348.

229

CHAPTER 6

PERCEPTUAL/PSYCHOMOTOR COMPUTER-ADMINISTERED MEASURES: FIELD TEST

Jeffrey J. McHenry, Jody L. Toquam, Rodney L. Rosse,
Norman G. Peterson, and Matthew K. McGue

In this chapter we describe analyses of the field test of the perceptual/psychomotor computer-administered measures in the Pilot Trial Battery, administered at Fort Knox in September 1984. The procedures and sample for this field test were described in Chapter 2, and the development and pilot testing of the computer-administered portion of the battery were described in Chapter 5. We note here that portions of this chapter are drawn from McHenry and McGue (1985) and Toquam, et al. (1985).

We present descriptions of the tests and discuss scoring issues and decisions. Descriptive statistics, reliability estimates, and uniqueness estimates for dependent measures or test scores are shown. The analyses of effects of video-game experience, computer testing station and practice on test scores are presented. Finally, the covariance of computer-administered test scores with each other, with the cognitive paper-and-pencil measures in the Pilot Trial Battery, and with ASVAB scores are presented.

PERCEPTUAL/PSYCHOMOTOR COMPUTERIZED TESTS ADMINISTERED

A concise description of each of the computer-administered tests included in the Pilot Trial Battery, along with a sample item or items from each test, is contained in Figure 6.1. Copies of the full Pilot Trial Battery administered at Fort Knox are contained in Appendix G. As Figure 6.1 shows, there are ten computer-administered tests in the Pilot Trial Battery, and these tests were intended to measure six constructs: Reaction Time, Perceptual Speed and Accuracy, Memory, Movement Judgment, Precision/Steadiness, and Multilimb Coordination.

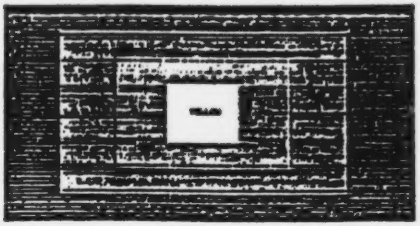
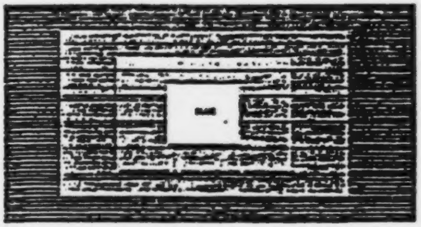
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
REACTION TIME		
Simple Reaction Time (Reaction Time Test 1)	The subject is instructed to place his/her hands on the green "home" buttons or in the Ready position. When the subject's hands are in the Ready position, a small box appears on the screen. After a delay period which varies from 1.5 to 3.0 seconds, the word YELLOW appears in the box. At this point, the subject must remove his/her preferred hand from the "home" buttons to strike the Yellow key on the testing panel. The subject must then return his/her hands to the ready position to receive the next item. The test contains 15 items. Although it is self-paced, subjects are given 10 seconds to respond before the computer time-outs and prepares to present the next item.	
Choice Reaction Time (Reaction Time Test 2)	Choice reaction time is assessed for two response alternatives only. This measure is obtained in virtually the same manner as the simple reaction time measure. The major difference involves stimulus presentation. Rather than presenting the same stimulus (YELLOW) on each trial, the stimulus varies. That is, subjects may see either of the stimuli BLUE or WHITE on the computer screen. When the stimulus appears, the subject is instructed to move his/her preferred hand from the "home" keys to strike the key that corresponds with the term (BLUE or WHITE) appearing on the screen. This test contains 15 items. Although the test is self-paced, the computer is programmed to allow the subject 9 seconds to respond before going on to the next item.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 1 of 8)

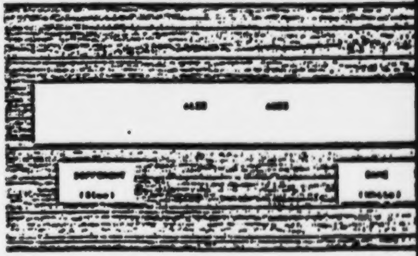
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
PERCEPTUAL SPEED AND ACCURACY		
Perceptual Speed and Accuracy Test	This test is designed to measure the ability to compare rapidly two visual stimuli presented simultaneously and determine whether they are the same or different. At the beginning of each trial, the subject is instructed to hold down the home keys. After a brief delay, the stimuli are presented. The subject must decide whether the stimuli are the same or different. He/she must then depress a white button if the stimuli are the same or a blue button if the stimuli are different. Four different "types" of stimuli are used: alpha, numeric, symbolic, and words. Within the alpha, numeric, and symbolic stimuli, the length of the stimulus is varied. Three different levels of length are presented: two-character, five-character, and nine-character. The test consists of 48 trials. The primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 2 of 8)

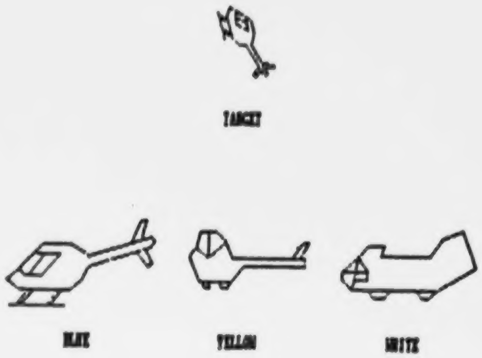
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
PERCEPTUAL SPEED AND ACCURACY (Continued)		
Target Identification Test	This test was designed to be a job-relevant measure of perceptual speed and accuracy. In this test, the subject is presented with a target object and three stimulus objects. The objects are pictures of military vehicles or aircraft (e.g., tanks, planes, helicopters), The target object is the same as one of the stimulus objects. However, the target may be rotated or reduced in size relative to its stimulus counterpart, or the target may be "moving" and growing across the screen. The subject must determine which of the three stimulus objects is the same as the target object and then press a button on the response pedestal corresponding to that choice. The test consists of 48 items; 24 are stationary, 24 are moving. The primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 3 of 8)

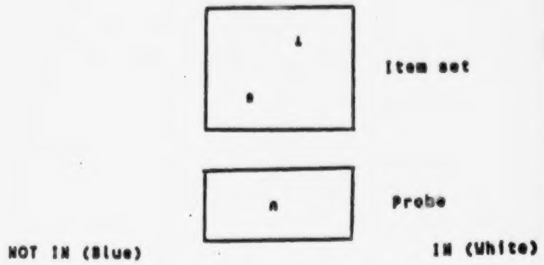
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
MEMORY		
Short-Term Memory Test	At the computer console, the subject is instructed to place his/her hands on the green home buttons. The first stimulus set then appears on the screen. A stimulus contains one, three, or five objects (letters or symbols). Following a delay period, the stimulus set disappears. When the probe appears, the subject must decide whether or not it was part of the stimulus set. If the probe was present in the stimulus set, the subject must strike the white key on the response pedestal. If the probe was not present, the subject must strike the blue key. The test includes 48 items. The primary dependent variable is the subject's average response time across those trials in which the subject makes a correct response.	
Number Memory Test	At the beginning of each trial of this test, the subject is presented with a single number on the computer screen. After studying the number, the subject is instructed to push a button to receive the next part of the problem. When the subject presses the button, the first part of the problem disappears and another number appears along with an operation term (e.g., "Add 9" or Subtract 6"). Once the subject has combined the first number with the second, he/she must press a button to receive a new number and operation term. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. In total, the test consists of 27 such items.	Start with 14 Divide by 7 Multiply by 8 NO (Blue) YES (White)

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 4 of 8)

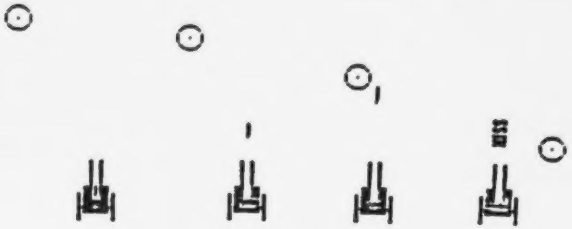
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
MOVEMENT JUDGMENT		
Cannon Shoot Test	At the beginning of each trial of this test, a stationary cannon appears on the computer console. The starting position of this cannon varies from trial to trial (i.e., it is positioned on the top, bottom, or side of the screen). The cannon is capable of firing a shell. The shell travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a response button to fire the shell such that the shell intersects the target when the target crosses the shell's line of fire. The test includes 48 items. The primary dependent variable is a deviation score indicating the difference between time of fire and optimal fire time (e.g., direct hits yield a deviation score of zero.)	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 5 of 8)

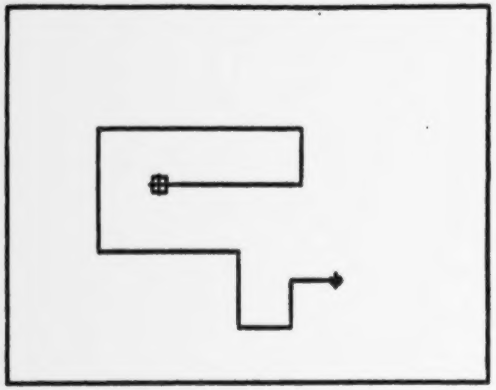
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
PRECISION/STEADINESS		
Target Tracking Test 1	This is a pursuit tracking test. On each trial of the test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box. Centered in the box is a crosshair. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject's task is to keep the crosshair centered within the target at all times. The subject uses a joystick to control movement of the crosshair. The subject's score on this test is the average distance from the center of the crosshair to the center of target across all 27 test trials.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 6 of 8)

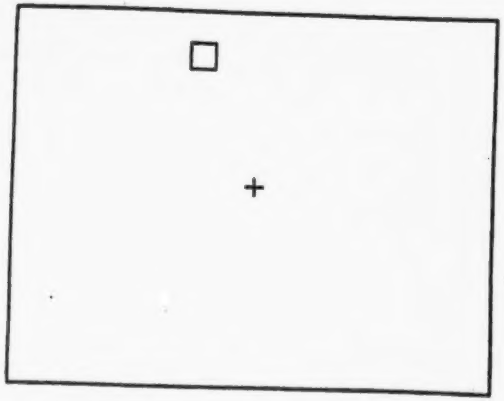
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
PRECISION/STEADINESS (continued)		
Target Shoot Test	At the beginning of a trial on this test, a crosshair appears in the center of the screen and a target box appears at some other location on the screen. The target then begins to move about the screen in an unpredictable manner, frequently changing speed and direction. The subject can control movement of the crosshair using a joystick. The subject's task is to move the crosshair into the center of the target. When this has been accomplished, the subject must press a red button on the response pedestal to "fire" at the target. The subject must do this before the time limit on each trial is reached. The subject receives three scores on this test. The first is the percentage of "hits" (i.e., the subject fires at the target when the crosshair is inside the target box). The second is the average time elapsed from the beginning of the trial until the subject fires at the target. The third score is the average distance from the center of the crosshair to the center of the target at the time the subject fires at the target. The test consists of 35 trials.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 7 of 8)

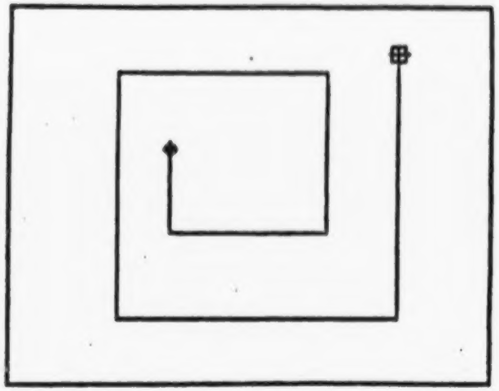
CONSTRUCT/MEASURE	DESCRIPTION OF TEST	SAMPLE ITEM
MULTILIMB COORDINATION		
Target Tracking Test 2	This is a test of multilimb coordination. The test is virtually identical to Target Tracking Test 1. The only difference is that the subject must use two sliding resistors (instead of a joystick) to control movement of the crosshair. The first sliding resistor controls movement of the crosshair in the vertical plane, while the second sliding resistor controls movement of the crosshair in the horizontal plane. As with Target Tracking #1, the subject's score on this test is the average distance from the center of the crosshair to the center of the target across all 27 test trials.	

Figure 6.1. Description of Perceptual/Psychomotor Computer-Administered Measures in Field Test. (Page 8 of 8)

Table 6.1 shows means, standard deviations, and reliability estimates for 19 scores or dependent measures for the 10 computer-administered tests. Before discussing this table and other aspects of the field test data analysis, we make a few remarks about the methods used to score these tests. In general, the methods employed were similar to those used at Fort Lewis (described in Chapter 5), but analyses of the Fort Knox field test data occasionally indicated a change was desirable.

Field Test Scoring Procedures

The perceptual computer-administered tests (see Table 6.1) generally yield one or both of two types of scores: accuracy and speed (except for the Cannon Shoot Test, discussed later)--for example, percent of items correct (accuracy) and mean reaction time (speed) on Perceptual Speed and Accuracy.

In addition, two derived measures can be computed for the perceptual tests: the slope and the intercept obtained when reaction times are regressed against an important defining characteristic of test items (which we called a "parameter"). For Perceptual Speed and Accuracy, this characteristic was the number of stimuli or characters being compared in an item (i.e., 2, 5, or 9 characters). In terms of speed of processing, the slope represents the average increase in reaction time with an increase of one character in the stimulus set; thus, the lower the value, the faster the comparison. The intercept represents all other processes not involved in comparing stimuli, such as encoding the stimuli and executing the response. Of course, these two measures can be used only when the test is well enough understood to allow the appropriate construction of items to tap a defining characteristic or parameter.

Reaction times on all tests were computed only for correct responses because it seemed to make very little sense to include incorrect responses. Subjects could simply respond at random and receive an excellent reaction time score if incorrect responses were included. This strategy means that items on most tests should be constructed so that subjects could answer every item correctly if given enough time, and that enough time is given. We did follow this strategy. Consequently, the speed measures (reaction time) were expected, in general, to have more variance and be more meaningful than the accuracy measures.

Several issues revolved around the choice of the particular way to measure reaction time. As noted in Chapter 5, total reaction time is made up of two components, decision time and movement time. Analyses of Fort Knox field test data indicated that total reaction time and decision time were very highly correlated and, since movement time is conceptually uninteresting, we elected to use total reaction time for all reaction time tests.

Means or medians across items could be used to compute the total reaction time scores. These could be trimmed (i.e., highest and lowest items not included in the calculation) or untrimmed (all items included). We looked at score distributions, intercorrelations of the various scores, and reliabilities of the scores in order to decide which method to use.

Table 6.1

Characteristics of the 19 Dependent Measures for Computer-Administered Tests: Fort Knox Field Tests (N = 256)[a]

Dependent Measure	Mean	SD	Reliability Split-Half (r_{sh})[b]	Reliability Test-Retest (r_{tt})[b]
PERCEPTUAL				
Simple Reaction Time (SRT)				
Mean Reaction Time (RT)	56.23 hs[c]	18.83 hs	.90	.37
Choice Reaction Time (CRT)				
Mean Reaction Time (RT)	67.41 hs	10.20 hs	.89	.56
Perceptual Speed and Accuracy (PS & A)				
Percent Correct (PC)	88%	8%	.83	.59
Mean Reaction Time (RT)	325.61 hs	70.38 hs	.96	.65
Slope	42.74 hs/ch[d]	15.56 hs/ch	.88	.67
Intercept	67.96 hs	45.02 hs	.74	.55
Target Identification				
Percent Correct (PC)	90%	10%	.84	.19
Mean Reaction Time (RT)	528.70 hs	133.96 hs	.96	.67
Short-Term Memory (STM)				
Percent Correct (PC)	85%	8%	.72	.34
Mean Reaction Time (RT)	129.68 hs	23.84 hs	.94	.78
Slope	7.22 hs/ch	4.53 hs/ch	.52	.47
Intercept	108.12 hs	23.18 hs	.84	.74
Number Memory				
Percent Correct (PC)	83%	13%	.63	.53
Mean Operation Time (RT)	230.71 hs	73.92 hs	.95	.88
Cannon Shoot				
Time Error (TE)	78.60 hs	20.28 hs	.88	.66
PSYCHOMOTOR				
Target Track 1				
Mean Log Distance	3.22	.44	.97	.68
Target Shoot				
Mean Time to Fire (std) (TF)	-.01	.48	.91	.48
Mean Log Distance (std)	-.01	.41	.86	.58
Target Track 2				
Mean Log Distance	3.91	.49	.97	.68

[a] N varies slightly from test to test.

[b] N = 120 for test-retest reliabilities, but varies slightly from test to test. r_{sh} = split-half reliability; odd-even item correlation with Spearman-Brown correction. r_{tt} = test-retest reliability, two week interval between administrations.

[c] hs = hundredths of a second

[d] hs/ch = hundredths of a second per character.

Generally, there were no striking differences between the methods. We decided to use untrimmed means for all tests except Simple and Choice Reaction Times; single extreme scores could affect the mean much more for these two tests than for the others because they had a much smaller number of items. Means were selected over medians because they had slightly higher reliabilities.

A final scoring issue concerns missing data. Since a subject may not get all items correct on a particular test, some information is missing when the mean total reaction time, slope, and intercept are being computed for that subject. Therefore, we established a maximum number of missing items that would be permitted for each test. This limit for all tests, with the exception of Number Memory, was set at 10 percent. Hence, for Simple and Choice Reaction Time, subjects could miss up to two items; for Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification, the limit was set at five items. Because Number Memory requires subjects to provide several responses for a single item, the possibility of missing data is higher. To ensure that sufficient numbers of subjects were available for analysis, we permitted subjects to miss up to seven of the 27 items in this test.

The Percent Correct and Mean Operation Time scores for the Number Memory Test require explanation since this test was not administered at Fort Lewis and, therefore, these scores were not discussed in Chapter 5. Percent Correct is simply the percentage of items that the subject answered correctly. Mean Operation Time is the mean of the mean reaction times to the four arithmetic operations (multiply, divide, add, and subtract). That is, for each subject, a mean reaction time for processing all the multiplication operations was computed; a separate mean for all the division operations, and so on for the two other operations. The mean of these four operation reaction time means was then computed and labeled Mean Operation Time.

As we noted above, procedures for scoring the Cannon Shoot Test differed from those used to score the other cognitive/perceptual tests. A reaction time score for this test is inappropriate because the task requires the subject to ascertain the optimal time to fire to ensure a direct hit on the target. (See description of Cannon Shoot Test, Figure 6.1.) Therefore, responses on this measure were scored by computing a deviation score that is composed of the difference between the time the subject fired and the optimal time to fire. These scores are summed across all items for each subject and a mean deviation time score is computed.

Scoring of two of the three psychomotor tests, Target Tracking Tests 1 and 2, was relatively straightforward. During each trial, the distance from the center of the crosshair to the center of the target was computed approximately 16 times per second, or almost 350 times per trial. These distances were then averaged by the computer, which outputs only the mean distance for each trial.

However, the frequency distribution of these mean distance scores proved to be highly positively skewed, the skewness coefficient for some trials being in excess of 5 and 6. Therefore, subjects' mean distance scores for each trial were transformed, using the natural logarithm transformation. The overall test score for each subject was then the mean of the log (mean distance) scores across the 27 trials of each test.

Scoring of the Target Shoot Test was a bit more complicated. Three overall test scores were generated for each subject: (1) the percentage of hits; (2) the mean distance from the center of the crosshair to the center of the target at the time of firing (the distance score); and (3) the mean time elapsed from the start of the trial until firing (the time-to-fire score). Percentage of hits was a less desirable measure because it contains relatively little information compared to the distance measure. Complications arose because subjects received no distance or time-to-fire scores on trials where they failed to fire at the target before the time limit for the trial elapsed. This scoring procedure resulted in considerable missing data; moreover, the missing data occurred primarily on the most difficult items of the test, where only the adept subjects were able to maneuver the crosshair close enough to the target to fire.

Therefore, as a first step in computing overall distance and time-to-fire scores for the Target Shoot Test, the distance and time-to-fire scores for each trial were standardized. That is, the mean and standard deviation of the distance score was computed for each item or trial on the test. Then, each subject was assigned a standard score on each trial by subtracting the item mean from his/her obtained distance score and dividing by the item standard deviation. For each subject, the overall distance and time score was then computed by averaging these standardized scores across all trials in which the subject fired at the target.

Mean Scores and Reliability Estimates

The means and standard deviations in Table 6.1 provide information about the score distributions. Note that the Percent Correct scores for Perceptual Speed and Accuracy, Target Identification, and Short-Term Memory are high, and the standard deviations are not large, as had been expected. The Reaction Time scores for these tests do have sufficient variance.

The split-half reliabilities range from .52 (Short-Term Memory Slope) to .96 (for two scores). Besides the Short-Term Memory Slope, only the Number Memory Percent Correct score is undesirably low (.63). All others are .74 or higher. These split-half reliabilities are odd-even correlations corrected to full test length, but note that they do not suffer from the artifactual inflation that speeded paper-and-pencil measures do. This is because all items are attempted by every subject.

The test-retest reliabilities are lower than the split-half reliabilities, as is typically the case. Three are so low as to cast doubt on the usefulness of the score: Simple Reaction Time Mean Reaction Time (.37), Target Identification Percent Correct (.19), and Short-Term Memory Percent Correct (.34). However, the two Percent Correct scores are not viewed as the primary score for their tests, and Simple Reaction Time is viewed largely as a "warm up" test. Although seven of the other scores have test-retest reliabilities below .60, there appears to be sufficient stability in these scores to warrant their possible use as predictors.

Uniqueness Estimates of Computer-Administered Test Scores

Table 6.2 shows uniqueness estimates for the 19 scores when regressed against the ASVAB subtests and the other computer-administered scores. The pattern of results here is similar to that found for the cognitive paper-and-pencil tests, except that the computer-administered tests have even higher U^2 coefficients, and thus show promise for adding to the validity obtained by the ASVAB. The exceptions are the Number Memory Scores. The two scores have lower uniqueness for ASVAB than for other computer tests. Several ASVAB subtests measure arithmetic and mathematical ability (Arithmetic Reasoning, Number Operations, and Mathematical Knowledge) and the Number Memory Test requires the use of the four basic arithmetic operations, so this finding, in retrospect, is not too surprising.

Later in this chapter we present the results of a factor analysis of the computer-administered test scores and the ASVAB sub-test scores which give additional information about the overlap between these two sets of tests.

Correlations with Video Game-Playing Experience

Table 6.3 shows correlations of the 19 computer-administered test scores with the subject's previous experience playing video games. In the computer-administered tests, the questions was asked: "In the last couple years, how much have you played video games on arcade machines, home video games or home computers?" Subjects selected one of the following five answers: "You have NEVER played video games," "You have tried a few games, but have generally played less than once a month," "You have played several times a month," "You have played at least once or twice a week," "You have played video games almost every day." These answers were given numeric values from 1 to 5, respectively. The mean score on this question was 2.99, SD = 1.03 (N = 256) and the test-retest reliability was .71 (N = 113).

Nine of the 19 correlations reached statistical significance at the .05 level, including three of the four scores from the psychomotor tests (Target Tracking 1 and 2 Mean Log Distances and Target Shoot Mean Log Distance). The Cannon Shoot score also showed a statistically significant correlation. Perceptual Speed and Accuracy, Target Identification, and Number Memory test scores showed no significant correlations, although Short-Term Memory did. The correlations are fairly low in general; the highest one is .27 with Target Shoot Mean Log Distance.

We interpret these findings as showing a small, but significant, relationship of video game-playing experience to the more "game-like" tests in the battery (i.e., the psychomotor tests), and a smaller, probably not meaningful, relationship with the cognitive/perceptual kinds of tests (with the possible exception of Short-Term Memory).

Effects of Differences in "Machine" or Computer Testing Station

We repeated the investigation which had been done at the pilot test at Fort Lewis on the effect of machine or computer testing station differences on computer-administered test scores. There were six computer testing stations in the field test, and approximately 40 male soldiers had been

Table 6.2

Uniqueness Estimates for the 19 Scores on Computer-Administered Tests in the Pilot Trial Battery Against Other Computer Scores and Against ASVAB

Score	Reliability		ASVAB		Other Computer Tests[a]	
	Split-Half $(r_{sh})^b$	Test-Retest $(r_{tt})^b$	R^2 With ASVAB[b]	U^{2c}	R^2 With Computer Scores[b]	U^{2c}
Simple Reaction Time						
Mean Reaction Time	.90	.37	.07	.83	.35	.55
Choice Reaction Time						
Mean Reaction Time	.89	.56	.09	.80	.44	.45
Perceptual Speed and Accuracy						
Percent Correct	.83	.59	.14	.69	.42	.41
Mean Reaction Time	.96	.65	.06	.90	.40	.56
Slope	.88	.67	.09	.79	.29	.59
Intercept	.74	.55	.11	.63	.19	.55
Target Identification						
Percent Correct	.84	.19	.05	.79	.25	.59
Mean Reaction Time	.96	.67	.16	.80	.64	.33
Short-Term Memory						
Percent Correct	.72	.34	.10	.62	.38	.34
Mean Reaction Time	.94	.78	.06	.88	.36	.58
Slope	.52	.47	.01	.51	.17	.35
Intercept	.84	.74	.11	.73	.34	.50
Number Memory						
Percent Correct	.63	.53	.40	.23	.18	.45
Mean Operation Time	.95	.88	.33	.62	.12	.83
Cannon Shoot						
Time Error	.88	.66	.02	.86	.12	.76
Target Track 1						
Mean Log Distance	.97	.68	.23	.74	.69	.28
Target Shoot						
Mean Time to Fire	.91	.48	.06	.85	.10	.81
Mean Log Distance	.86	.58	.11	.75	.33	.53
Target Track 2						
Mean Log Distance	.97	.77	.17	.80	.67	.30

[a] In computing the R^2 with other computer tests, each test score was predicted using only the test scores from the remaining nine computer tests. Thus, for example, STM-Intercept was not used as a predictor in estimating STM-Mean RT.

[b] The R^2 with the ASVAB and with the other computer-administered tests were corrected for shrinkage that would be expected with cross-validation. N = 182 for R^2 computations.

[c] Uniqueness estimates (U^2) were computed using the split-half reliability estimate. The uniqueness is equal to the reliability minus the R^2 with the ASVAB or with the other computer tests. It is a measure of the unique, reliable variance that each test score might contribute to the prediction of job performance criteria.

Table 6.3

Correlations Between Computer Test Scores and Previous Experience With Video Games (N = 250)[a]

Computer Test	Test Score	Correlation[b]
Simple Reaction Time	Mean RT	.12*
Choice Reaction Time	Mean RT	.15*
Perceptual Speed & Accuracy	Percent Correct	-.01
	Mean RT	.01
	Slope	-.03
	Intercept	.06
Target Identification	Percent Correct	.08
	Mean RT	.05
Short-Term Memory	Percent Correct	.13*
	Mean RT	.08
	Slope	-.16*
	Intercept	.18*
Number Memory	Percent Correct	.08
	Mean RT	.00
Cannon Shoot	Time Error	.18*
Target Tracking 1	Mean Log Distance	.22*
Target Shoot	Mean Time to Fire	.10
	Mean Log Distance	.27*
Target Tracking 2	Mean Log Distance	.16*

---

[a] Varies slightly by test.

[b] Correlations of .12 or greater are statistically significant at the .05 level, two-tailed test of significance. Signs of correlations have been reflected, where appropriate, so that greater video experience shows positive correlation with better test performance.

tested at each station. (We used only males in this analysis to avoid confounding the results with gender differences, since the 47 females tested were not evenly balanced across the six testing stations. Also, only males with complete sets of computer test scores were used so the analyses would have the same sample for each test score.)

We ran a one-way multivariate analysis of variance (MANOVA) for the 19 computer test scores, with six "machine" levels. As Table 6.4 shows, machine differences had no effect on test scores. The MANOVA likelihood ratio was .99 (p value = .50). Table 6.4 also shows the univariate F ratio and p values for each of the 19 scores. None of them reached statistical significance at the .05 level, again indicating that the testing station had no significant effect on these 19 scores.

These results were especially encouraging because they replicated a similar set of results from the earlier Fort Lewis pilot test (see Chapter 5). The results showed that the hardware and software used in the computer-administered battery had, indeed, resulted in a standardized testing situation across the six machines and testing stations. We think this is due in large part to the calibration software used to make the hardware equivalent across stations, as described in Chapter 1.

Table 6.4

Effects of Machine Differences on Computer Test Scores[a]: Fort Knox Field Test

Computer Test Score	F	p[b]
Simple Reaction Time		
Mean Reaction Time	1.59	.16
Choice Reaction Time		
Mean Reaction Time	.52	.76
Perceptual Speed and Accuracy		
Percent Correct	1.18	.32
Mean Reaction Time	.56	.73
Slope	.84	.53
Intercept	.85	.52
Target Identification		
Percent Correct	1.67	.14
Mean Reaction Time	.93	.46
Short-Term Memory		
Percent Correct	.11	.99
Mean Reaction Time	.95	.45
Slope	1.13	.34
Intercept	.64	.67
Number Memory		
Percent Correct	.56	.73
Mean Operation Time	1.55	.17
Cannon Shoot		
Time Error	2.14	.06
Target Track 1		
Mean Log Distance	.62	.69
Target Shoot		
Mean Time to Fire	1.91	.09
Mean Log Distance	1.01	.41
Target Track 2		
Mean Log Distance	.86	.51

[a] MANOVA likelihood ratio = .99, p = .50 for these test scores.

[b] Degrees of freedom (df) = 5,200 for all 19 test scores.

EFFECTS OF PRACTICE ON SELECTED COMPUTER-ADMINISTERED TEST SCORES

During the Fort Knox field test, data were collected to investigate the effects of practice on computer test scores. The experimental design for this work is shown in Figure 6.2. In accordance with this design, a statistically significant Time x Group interaction would indicate that a practice effect had occurred.

Figure 6.3 shows the make-up of the test items in the computer practice battery and the order in which they were administered. Practice was given on five tests: Reaction Time 2 (Choice Reaction Time), Target Tracking 1, Cannon Shoot, Target Tracking 2, and Target Shoot. These tests were selected because they were thought to be the tests that would show greatest improvement with practice. All the psychomotor tests were included. The soldiers in the practice group received two practice sessions on each of the five tests and then completed the five tests as they had been administered to them the first time they completed the battery. Note that unique items (i.e., items not appearing on the full battery test) were used for Target Tracking 1, Target Tracking 2, and Cannon Shoot.

Table 6.5 shows the results of the ANOVAs for the five tests included in the practice effects research. (We initially used separate ANOVAs rather than a MANOVA, knowing that it could spuriously show significant effects where a MANOVA would not. However, when only one practice effect reached statistical significance, it seemed unnecessary to run the more conservative MANOVA.) These results show only one statistically significant practice effect, the Mean Log Distance score on Target Tracking 2. Three findings for Time were statistically significant, indicating that scores did change with a second testing, whether or not practice trials intervened between the two tests. Finally, note that the omega-squared values show that relatively small amounts of test score variance are accounted for by the Group, Time or Time x Group factors, also demonstrating the insignificance of practice effects.

Table 6.6 shows further analyses of the practice experimental data. Gain scores and test-retest reliability coefficients were computed for the retest and practice groups, and tests for significant differences between the two groups were performed. Note that the difference between the gain scores for the retest and practice groups reached statistical significance only for the distance score for Target Tracking 2, reflecting the same finding in Table 6.5.

These data suggest that the practice intervention was not a particularly strong one. It should be noted, though, that on some tests subjects' performance actually deteriorated from Time 1 to Time 2. The average gain score for the two groups across the five dependent measures was only .09 standard deviations. This suggests either that the tasks used in these tests are resistant to practice effects, or that performance on these tasks reaches a maximum level of proficiency after only a few trials. Also, recall that analyses of the PTB cognitive paper-and-pencil tests (see Table 4.3) showed gain scores that were as high as or higher than those found here. Perhaps gain in scores through retesting or practice is of even less concern for computerized tests than for paper-and-pencil tests.

Group 1 T_1 ----------------------------> T_2 Retest Group
Two Weeks (N = 113)

Group 2 T_1 --------------> Practice ----->T_2 Practice Group
Two Weeks (Five Tests) (N = 74)

ANOVA

Source	DF
Group	A-1
Subjects (Group)	(B-1)A
Time	C-1
Time x Group	(C-1)(A-1)
Time x Subject (Group)	(C-1)(B-1)A

Practice Effect = Significant Time x Group Interaction

Figure 6.2 Experimental design of the practice effects investigation.

Test	No. of Items	Comments
Demographics	5	Same as in the Test Battery
Reaction Time 2	15	Same as in the Test Battery
Target Tracking 2	15	Unique items
Cannon Shoot	24	Unique items
Target Tracking 2	15	Unique items
Target Shoot	20	Odd-numbered items from the Test Battery
Reaction Time 2	15	Same as in the Test Battery
Target Tracking 1	15	Unique items
Cannon Shoot	24	Unique items
Target Tracking 2	15	Unique items
Target Shoot	20	Odd-numbered items from the Test Battery
Reaction Time 2	15	Same as in the Test Battery
Target Tracking 1	27	Same as in the Test Battery
Cannon Shoot	48	Same as in the Test Battery
Target Tracking 2	27	Same as in the Test Battery
Target Shoot	40	Same as in the Test Battery

Figure 6.3 Items in the Computer Practice Battery used at the Fort Knox Field Test.

Table 6.5

Effects of Practice on Selected Computer Test Scores

Test	Dependent Measure	Source of Variance	df	F	Omega Squared
Choice Reaction Time	Trimmed Mean Reaction Time	Group	1,180	9.71*	.032
		Time	1,180	25.70*	.035
		Time x Group	1,180	.73	--
Target Tracking 1	Mean Log Distance	Group	1,178	.73	--
		Time	1,178	9.26*	.005
		Time x Group	1,178	4.11	--
Target Tracking 2	Mean Log Distance	Group	1,178	.47	--
		Time	1,178	1.30	--
		Time x Group	1,178	7.79*	.005
Cannon Shoot	Time Error	Group	1,171	3.79	--
		Time	1,171	.16	--
		Time x Group	1,171	5.72	--
Target Shoot	Mean Log Distance	Group	1,171	.41	--
		Time	1,171	9.28*	.012
		Time x Group	1,171	.08	--

*Denotes significance at $p<.01$.

Next, Table 6.6 shows that the test-retest stability for all five dependent measures was greater for the retest group than for the practice group. (While the difference between the stability coefficients for the two groups was statistically significant for only one of the dependent measures, the test was not very powerful; statistical significance required a difference of approximately .40 between the two stabilities.) Closer inspection of the data shows that the stability coefficients for the two groups were very nearly equal for the three "distance" dependent measures. Thus, it appears that the rank-ordering of subjects' performance on psychomotor tests is not greatly affected by practice.

Another method for examining practice effects is to look at the correlations between items or parts within a test. This was done for Target Tracking Tests 1 and 2. Each test was divided into three parts corresponding to test items 1-9, 10-18, and 19-27. A distance score was then computed for each of the three parts. Table 6.7 shows the intercorrelations among the three part scores for both tests for both Time 1 and Time 2. (Time 2 data were taken from the retest group only; the practice group's data were not included.)

If the ability requirements of the tracking task were changing due to practice during the course of the test, one would expect to find that the correlation between items 1-9 and items 19-27 would be lower than either of the two correlations involving items 10-18. This did not occur. While

Table 6.6

Gain Scores and Reliabilities for Retest and Practice Groups[a]

Test	Dependent Measure	Group	Gain Score[b]	*F* for Gain Scores	Reliability	*Z* for Reliability[c]
Choice Reaction Time	Trimmed Mean Reaction Time	Retest	-.36	.73	.56	1.64
		Practice	-.43		.36	
Target Tracking 1	Mean Log Distance	Retest	.07	4.11	.68	.46
		Practice	.33		.64	
Target Tracking 2	Mean Log Distance	Retest	-.09	7.79*	.77	.16
		Practice	.21		.76	
Cannon Shoot	Time Error	Retest	.34	5.72	.66	1.50
		Practice	-.11		.51	
Target Shoot	Mean Log Distance	Retest	.21	.08	.58	.88
		Practice	.26		.48	

[a] Inferential statistics significant at $p < .01$ are denoted with an asterisk(*).

[b] Gain scores are effect size estimates and were computed using the pooled standard deviation. Signs were reflected as necessary so that a positive gain score denotes "improvement" from Time 1 to Time 2.

[c] Given the sizes of the retest and practice samples, statistical significance (at $p < .01$) will not be attained until the difference between the two reliabilities reaches approximately .40.

Table 6.7

Intercorrelations Among Items 1-9, Items 10-18, and Items 19-27 of Target Tracking Tests 1 and 2

Target Tracking Test 1							
Time 1				Time 2			
	Items 1-9	Items 10-18	Items 19-27		Items 1-9	Items 10-18	Items 19-27
Items 1-9				Items 1-9			
Items 10-18	.87			Items 10-18	.91		
Items 19-27	.80	.87		Items 19-27	.92	.92	

Target Tracking Test 2							
Time 1				Time 2			
	Items 1-9	Items 10-18	Items 19-27		Items 1-9	Items 10-18	Items 19-27
Items 1-9				Items 1-9			
Items 10-18	.83			Items 10-18	.86		
Items 19-27	.85	.89		Items 19-27	.85	.91	

there is a slight tendency for the correlation between items 10-18 and items 19-27 to be the highest of the three intercorrelations, the difference between the highest and lowest correlation within each test averages only .05. Data in Table 6.1 show that the Spearman-Brown corrected split-half reliability of both tests is .97, suggesting that all of the items within each test are measuring the same underlying ability.

In summary, data from the practice experiment indicate that scores from computerized psychomotor tests appear to be quite stable over a two-week period. Practice does have some effect on test scores, but it appears to be relatively small. Certainly it does not seem strong enough to warrant serious concern about the usefulness of the tests.

COVARIANCE ANALYSES WITH ASVAB SUBTESTS AND COGNITIVE PAPER-AND-PENCIL TESTS

Table 6.8 contains the intercorrelations for the ASVAB subtests, paper-and-pencil cognitive measures, and the computer-administered tests, which include both perceptual and psychomotor measures. Scores on the AFQT are also included. These correlations are based on the Fort Knox field test sample but include only those subjects with test scores available on all variables (N = 168).

In examining these relationships, we first looked at the correlations between tests within the same battery. As was discussed in Chapter 4, correlations between ASVAB subtest scores range from .02 to .74 (absolute values), and correlations between the cognitive paper-and-pencil test scores range from .27 to .67. For the perceptual computer-administered test scores, correlations range from .00 to .83 (absolute terms). Note that the highest values appear for correlations between scores computed from the same test; for example, the correlation between Short-Term Memory reaction time and intercept is .83, and the correlation between Perceptual Speed and Accuracy slope and reaction time is .82. Correlations between the psychomotor computer-administered variables range from .15 to .81 (absolute terms). Note that scores on the two tracking tests correlate the highest.

Perhaps the most important question to consider is the overlap between the different groups of measures. Do the paper-and-pencil measures and computer-administered tests correlate highly with the ASVAB and with each other or are they measuring unique or different abilities? To address this question, in part, we examined the intercorrelations between the ASVAB, including AFQT, and other groups of tests.

As noted in Chapter 4, for the cognitive paper-and-pencil tests these correlations range from .01 (Assembling Objects and Number Operations) to .63 (Orientation 3 and Mechanical Comprehension), with a mean correlation of .33 (see Table 6.9 for a summary of the correlation statistics). Across all PTB paper-and-pencil tests, ASVAB Mechanical Comprehension appears to correlate the highest with the new tests; across all ASVAB subtests, PTB Orientation 3 yields the highest correlations.

The correlations between the ASVAB subtests and the computer-administered perceptual tests, in absolute terms, range from .00 (Paragraph Comprehension with Perceptual Speed and Accuracy Reaction Time and with Short-Term Memory Intercept, and General Science with Perceptual Speed and Accuracy Slope) to .58 (Arithmetic Reasoning and Number Memory Percent Correct). The mean of these 165 correlations is .15 (SD = .12). Across all ASVAB subtests, scores on the Short-Term Memory Reaction Time and Slope yield the lowest correlations. The highest values appear for Number Memory Percent Correct and Reaction Time.

The correlations between ASVAB subtests and psychomotor scores range from .00 (Coding Speed with Target Shoot Time and Target Shoot Distance) to -.44 (Mechanical Comprehension and Tracking 1). The mean of these 44 correlations (absolute values) is .17 (SD = .12). Note that for the most part, these four PTB variables yield the highest correlations with ASVAB Mechanical Comprehension and Electronics Information. The lowest correlations appear for Paragraph Comprehension, Number Operations, and Coding Speed.

Table 6.8

Intercorrelations Among the ASVAB Subtests and the Pilot Trial Battery Cognitive Paper-and-Pencil and Perceptual/Psychomotor Computer-Administered Tests: Fort Knox Sample

(N = 168)

		AFQT Score	Gen Scienc	Arith Reas	Word Know	Parg Comp	Numb Ops	Code Spd	Auto/Shop	Math Know	Mech Comp	Elec Info	Asmbl Obj	Obj Rotat	Shapes	Maze	Path	Reas 1	Reas 2	Orient 1	Orient 2	Orient 3	SRT-RT	CRT-RT	PSSA-PC	PSSA-RT	PSSA-Slp	PSSA-Int	Targ ID-PC	Targ ID-RT	STM-PC	STM-RT	STM-Slp	STM-Int	Cum Shoot	No Mem-PC	No Mem-RT	Trk 1-Dist	Trk 2-Dist	TSht-Time	TSht-Dist
ASVAB	AFQT Score																																								
	Gen Scienc	61																																							
	Arith Reas	87	56																																						
	Word Know	81	66	61																																					
	Parg Comp	69	43	53	58																																				
	Numb Ops	44	02	21	06	14																																			
	Code Spd	30	-02	14	10	14	53																																		
	Auto/Shop	45	34	50	43	29	-08	-07																																	
	Math Know	76	60	76	62	54	19	15	43																																
	Mech Comp	55	50	62	54	36	-10	-06	64	57																															
	Elec Info	56	66	56	55	39	-03	01	71	59	63																														
PTB - Paper-and-Pencil Tests	Asmbl Obj	44	38	48	48	29	-01	19	39	48	57	38																													
	Obj Rotat	30	20	33	18	18	18	13	27	29	36	32	47																												
	Shapes	35	29	33	28	20	16	23	17	36	35	26	51	50																											
	Maze	32	23	38	16	12	19	19	36	36	44	35	60	58	47																										
	Path	33	20	37	21	17	24	31	33	35	42	32	54	59	50	63																									
	Reas 1	46	33	48	41	33	04	15	29	47	51	32	66	38	41	51	52																								
	Reas 2	41	32	44	36	36	02	12	18	49	38	30	52	27	44	33	37	50																							
	Orient 1	45	41	45	35	29	14	20	40	47	50	39	52	55	54	55	58	49	41																						
	Orient 2	50	29	52	38	33	18	13	36	48	49	35	51	38	42	49	42	48	47	55																					
	Orient 3	59	51	62	48	43	05	15	55	60	63	54	61	42	42	51	46	49	53	67	56																				
PTB - Computerland Perceptual Tests	SRT-RT	-23	-11	-20	-15	-15	-25	-07	-09	-14	-11	-08	-11	-21	-09	-22	-23	-29	-16	-17	-31	-18																			
	CRT-RT	-16	-05	-21	-02	-08	-20	-14	-18	-10	-14	-16	-16	-37	-19	-37	-33	-21	-12	-27	-37	-22	53																		
	PSSA-PC	25	-01	22	18	23	14	30	04	12	11	02	18	16	10	02	19	25	14	13	10	16	17	17																	
	PSSA-RT	-06	-04	-01	-02	00	-14	-15	-09	-06	-12	-18	-15	-17	-24	-36	-18	-03	01	-13	-16	-13	19	31	50																
	PSSA-Slp	11	-00	17	07	09	-02	-01	-05	06	01	-14	06	-07	-12	-15	-01	12	16	00	01	-01	-03	09	52	82															
	PSSA-Int	-23	-03	-30	-10	-12	-13	-15	-06	-16	-17	-01	-31	-21	-16	-28	-28	-30	-25	-20	-26	-15	32	31	-27	-08	-61														
	Targ ID-PC	12	07	07	21	09	-08	10	07	06	12	01	12	05	09	-05	13	12	14	05	03	11	11	07	40	33	32	-13													
	Targ ID-RT	-23	-31	-29	-19	-06	-05	-02	-34	-25	-37	-32	-41	-46	-34	-50	-45	-29	-20	-34	-37	-35	23	42	20	47	32	13	16												
	STM-PC	21	05	16	14	12	17	25	01	05	12	05	15	10	15	02	18	14	23	10	20	20	-06	06	50	17	26	-23	25	03											
	STM-RT	-02	02	-10	11	05	-09	-13	-02	-01	-09	-08	-13	-20	-13	-28	-22	-06	-05	-12	-20	-06	23	40	26	49	25	23	27	47	08										
	STM-Slp	12	06	09	07	10	10	06	01	08	11	-02	09	-00	05	02	18	10	09	05	06	11	-06	03	28	29	26	-11	18	13	32	39									
	STM-Int	-10	-02	-17	08	-00	-16	-17	-03	-06	-16	-08	-19	-22	-16	-31	-29	-12	-11	-17	-25	-16	28	42	10	35	11	31	18	42	-11	03	-19								
	Cum Shoot	-11	-13	-16	-13	01	-06	-03	-17	-03	-15	-15	-16	-17	-07	-15	-17	-08	-03	-11	-08	-08	13	10	00	08	-01	11	-09	25	-02	25	14	10							
	No Mem-PC	60	42	58	41	36	35	20	33	56	35	34	35	26	25	29	32	40	31	37	33	43	-16	-09	29	02	15	-20	16	-12	23	-00	02	-01	-18						
	No Mem-RT	-49	-27	-46	-22	-21	-45	-29	-23	-42	-23	-31	-17	-26	-25	-36	-26	-20	-18	-35	-29	-36	21	24	11	36	21	03	10	27	07	18	15	11	08	-45					
PTB - Computerland Psychomotor Tests	Trk 1-Dist	-20	-21	-31	-16	-07	09	-06	-33	-24	-44	-28	-38	-34	-23	-35	-41	-29	-26	-28	-39	-36	14	25	-12	08	00	11	-06	42	-29	25	-15	35	27	-16	-08				
	Trk 2-Dist	-19	-23	-28	-15	-03	07	-05	-33	-21	-38	-29	-37	-28	-14	-32	-35	-22	-22	-25	-36	-32	11	19	-01	11	06	09	02	39	-19	25	-01	27	30	-14	02	81			
	TSht-Time	-09	-08	-16	-03	05	-02	-08	-18	-10	-21	-17	-10	-14	-02	-23	-17	-11	-01	-09	-14	-06	08	16	16	22	09	12	12	32	09	22	15	15	12	-10	15	23	19		
	TSht-Dist	-09	-19	-17	-09	-02	13	00	-26	-12	-31	-23	-31	-22	-14	-30	-37	-26	-14	-23	-26	-28	08	16	-07	03	-00	09	-11	32	-12	27	-16	38	25	-08	02	60	55	-13	

The intercorrelations between the PTB cognitive paper-and-pencil tests and the computerized tests in general range from .00 to .46 (in absolute terms). The mean of the 40 psychomotor/cognitive paper-and-pencil test score correlations is .24 (SD = .11). The mean of the 150 perceptual computer score/cognitive paper-and-pencil test score correlations is .19 (SD = .1). The computerized test variables that correlate consistently highly with the paper-and-pencil tests include Target Identification Reaction Time, Number Memory Percent Correct and Reaction Time, Tracking 1, and Tracking 2.

Intercorrelations between the cognitive/perceptual computer tests and the psychomotor computer tests range from .00 to .42 (mean = .15 and SD = .11). The highest values appear for the correlations between the four psychomotor measures and Target Identification Percent Correct and Short-Term Memory Slope.

Table 6.9 summarizes the correlational data in Table 6.8 that we discussed just above. The values in the two tables and the discussion lead to the conclusion that the various types of measures do not overlap excessively, and, therefore, do appear to each make separate contributions to ability measurement.

Table 6.9

Mean Correlations, Standard Deviations, and Minimum and Maximum Correlations Between Scores on ASVAB Subtests and Pilot Trial Battery Tests of Cognitive, Perceptual, and Psychomotor Abilities

Types of Scores Correlated	Number of Correlations	Mean[a] Correlations	SD[a] of Correlation	Minimum[a] Correlation
ASVAB Subtests and PTB Cognitive Paper-and-Pencil Tests	110	.33	.14	.01
ASVAB Subtests and PTB Cognitive/ Perceptual Computer-Administered Tests	165	.15	.12	.00
ASVAB Subtests and PTB Psychomotor Computer-Administered Tests	44	.17	.12	.00
PTB Cognitive Paper-and-Pencil Tests and PTB Perceptual Computer-Administered Tests	150	.19	.11	.00
PTB Cognitive Paper-and-Pencil Tests and PTB Psychomotor Computer-Administered Tests	40	.24	.11	.01
PTB Perceptual Computer-Administered Tests and PTB Psychomotor Computer-Administered Tests	60	.15	.11	.00

[a] These statistics are based on absolute correlation values.

FACTOR ANALYSIS OF PTB COGNITIVE PAPER-AND-PENCIL MEASURES, PTB PERCEPTUAL-PSYCHOMOTOR COMPUTER-ADMINISTERED TESTS, AND ASVAB SUBTESTS

In addition to examining intercorrelations, we also examined results from a factor analysis of scores of the ASVAB, cognitive paper-and-pencil measures, and computer-administered tests. Two variables, Perceptual Speed and Accuracy Reaction Time and Short-Term Memory Reaction Time, were omitted from this analysis because these scores correlated very highly with their corresponding Slope or Intercept variables; to avoid obtaining communalities greater than one, these two reaction time measures were omitted.

Results from the seven-factor solution of a principal components factor analysis with varimax rotation are displayed in Table 6.10. All loadings of .30 or greater are shown. Our interpretation of these data, by factor, is as follows.

- Factor 1 includes eight of the ASVAB subtests (General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Automotive Shop, Mathematical Knowledge, Mechanical Comprehension, and Electionics Information), six of the cognitive paper-and-pencil measures (Assembling Objects, Reasoning 1 and 2, and Orientation 1, 2, and 3) and two perceptual computer variables (Number Memory Percent Correct and Reaction Time). Because this factor contains measures of verbal, numerical, and reasoning ability we have termed this "g", or a general ability factor.

- Factor 2 includes all of the PTB cognitive paper-and-pencil measures, Mechanical Comprehension from the ASVAB, and Target Identification Reaction Time from the computer tests. We called this a general spatial factor.

- Factor 3 has major loadings on the three psychomotor tests (Tracking 1, Tracking 2, and Target Shoot Distance), with substantially smaller loadings from three cognitive/perceptual computer test variables (Target Identification Reaction Time, Short-Term Memory Intercept, and Cannon Shoot Time Error), the Path Test, and Mechanical Comprehension from the ASVAB. Given the high loadings of the psychomotor tests on this factor, we refer to this as the motor factor.

- Factor 4 includes variables from the cognitive/perceptual computer tests. These include PS&A Percent Correct, Slope, and Intercept; Target Identification Percent Correct, and Short-Term Memory Percent Correct. This factor appears to involve accuracy of perception across several tasks and types of stimuli.

- Factor 5 contains variables from the perceptual computer tests, including Simple Reaction Time RT, Choice Reaction Time RT, Short-Term Memory Intercept, PS&A Intercept and Percent Correct, and Target ID RT. Also loading on this factor is a cognitive paper-and-pencil test, Orientation 2. This factor is not very clear, but the highest loadings are on straightforward reaction time measures, so we interpret this as a speed of reaction factor.

Table 6.10

Principal Components Factor Analysis of Scores of the ASVAB Subtests, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer-Administered Tests[a]

(N = 168)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	h^2
ASVAB								
GS	75							59
AR	75							73
WK	77							62
PC	62							47
NO						84		77
CS						62		44
AS	62							58
MK	77							70
MC	63	38	-30					68
EI	72							65
COGNITIVE PAPER-AND-PENCIL								
Assemb Obj	35	69						66
Obj Rotation		-61						49
Shapes		66						51
Maze		70						67
Path		67	-30					65
Reason 1	37	58						54
Reason 2	37	47						44
Orient 1	37	64						58
Orient 2	40	46			-30			52
Orient 3	60	52						67
PERCEPTUAL COMPUTER								
SRT-RT					63			44
CRT-RT					61			50
PS&A-PC				67	31			70
PS&A Slope				88				81
PS&A Inter				-65	50			74
Target ID-PC				40				25
Target ID-RT		-41	37		30			57
STM-PC				39			34	41
STM-Slope							41	25
STM-Int			38		51			47
Cannon Shoot-TE			32					19
No Mem-PC	53					37		52
No Mem-RT	-37					-46		54
PSYCHOMOTOR COMPUTER								
Tracking 1			86					82
Tracking 2			77					66
Target Shoot-TF							42	23
Target Shoot-Dist			64					48
Variance Explained	5.69	4.70	2.83	2.37	1.92	1.87	1.17	

NOTE: Decimals have been omitted from factor loadings.

[a] Note that the following variables were not included in this factor analysis: AFQT, PS&A, Reaction Time, and Short-Term Memory Reaction Time.

h^2 = communality (sum of squared factor loadings) for variables.

- Factor 6 contains four variables, two from the ASVAB (Number Operations and Coding Speed) and two from the perceptual computer tests (Number Memory Percent Correct and Reaction Time). This factor appears to represent both speed of reaction and arithmetic ability.

- Factor 7 contains three variables from the computer-administered tests: Short-Term Memory Percent Correct and Slope, and Target Shoot Time to Fire. This factor is difficult to interpret, but we believe it may represent a response style factor. That is, this factor suggests that those individuals who take a longer time to fire on the Target Shoot Test also tend to have higher slopes on the Short-Term Memory Test (lower processing speeds with increased bits of information) but are more accurate or obtain higher percent correct values on Short-Term Memory.

Note that several variables--Target Identification Percent Correct, Short-Term Memory Percent Correct, Cannon Shoot Time Error, and Target Shoot Time to Fire--have fairly low communalities. These may be due to relatively low score variance or reliability, but it could also be due to those variables having unique variance, at least when factor analyzed with this set of tests. We think this latter explanation is highly plausible for the Cannon Shoot score.

This concludes the discussion of the pilot testing and the Fort Knox field test of the cognitive paper-and-pencil tests and the computer-administered tests in the Pilot Trial Battery. We turn now to a discussion of the non-cognitive measures in Chapters 7 and 8.

Chapter 6 References

McHenry, J. J., & McGue, M. K. (1985). *Problems, issues, and results in the development of computerized psychomotor measures.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hanson, M. A. (1985). *Development of cognitive/perceptual measures: Supplementing the ASVAB.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

CHAPTER 7

NON-COGNITIVE MEASURES: PILOT TESTING

Leaetta M. Hough, Bruce N. Barge, and John D. Kamp

GENERAL

In this chapter, we describe the development and pilot testing of the non-cognitive measures prepared for inclusion in the Pilot Trial Battery. All are paper-and-pencil measures. The inventories developed tap constructs in the temperament, interest, and life history (biodata) domains. Field testing of these measures is covered in Chapter 8.

The non-cognitive measures were pilot tested at Fort Campbell and Fort Lewis in the spring of 1984. In addition to the newly developed measures, four published, marker measures of temperament were utilized in the pilot tests. Chapter 2 contains a detailed description of the pilot test procedures and samples and we do not repeat that discussion here. The pilot test results are discussed later in this chapter; we first discuss the desired characteristics of these measures.

Desired Characteristics

As described in Chapter 1, the Task 2 research team extensively reviewed the literature and the existing tests and constructs available in the non-cognitive area as well as in the cognitive and psychomotor areas. The literature review served to identify non-cognitive constructs most relevant and important for the prediction of success in a variety of Army MOS (Hough, Kamp, & Barge, 1985).

In the non-cognitive area, there was particular interest in predicting "adjustment" criteria, such as attrition, job satisfaction, and unfavorable discharge/disciplinary action, as well as job and training performance. Attention to adjustment criteria was important in the development of non-cognitive predictors because these criteria are typically not highly related to scores on cognitive or perceptual/psychomotor tests. Non-cognitive measures were also seen as valuable for use in classification. The expert judgment research (see Chapter 1) indicated the importance of including measures of several non-cognitive constructs. Following these explorations, the IPR meeting in March 1984 resulted in the identification of a set of non-cognitive constructs to be developed for the Pilot Trial Battery. (See Figure 1.5.)

Development of the non-cognitive measures was guided by several important, yet sometimes conflicting, goals. First, it was desired that the scales have construct validity. Item content of each scale should be heterogeneous enough to cover all important aspects of the targeted construct, yet homogeneous enough to be interpretable and distinct from other constructs. In addition, the scales should be a valid assessment of the respondent's standing on the construct, rather than merely a reflection of social desirability.

Other important considerations during the development of the inventories included reliability and stability. The scales were to be both internally consistent and stable over time (test-retest). The measures should also be stable over situations, so that faking or differing response sets would not greatly distort the scores obtained. Items and scales should elicit sufficient variance in responses that the scores could be used to differentiate respondents. It was important that the item content be non-objectionable. Finally, it was extremely important that the measures be able to demonstrate validity in predicting the respondent's standing on various job performance and other important criteria.

ABLE and AVOICE

The above set of desired characteristics formed the basis for the development of the scales to be described in this chapter. Our discussion of these scales is divided into two areas that correspond to the two inventories that were employed. The ABLE (Assessment of Background and Life Experiences) contains items that assess the important constructs of the temperament and life history (biodata) domains. The items on the ABLE are all new items written by PDRI researchers. Each item was written to tap one of the constructs identified via the literature review and other earlier phases of the project (see above and Chapter 1). Many candidate items were written. These were reviewed by the entire non-cognitive team and the best appearing items were selected for initial inclusion on the ABLE. The main criteria for item selection were: the item was clearly relevant for measuring a targeted construct; it was clearly written, and content was non-objectionable. The AVOICE (Army Vocational Interest Career Examination) measures the relevant constructs of the interest domain. The AVOICE is a significantly modified version of the VOICE (Vocational Interest Career Examination) which had been developed and researched by the U.S. Air Force (Alley & Matthews, 1982). In general, items were modified to measure interests that seemed more appropriate to Army occupations. Items were also written to tap interests that were not included on the VOICE. We describe the constructs, scales, and pilot test results of the ABLE first, and then do the same for the AVOICE.

The constructs chosen for the battery are described with examples of the item content for each construct scale; any revisions made on the basis of the pilot tests are discussed. Data obtained during the pilot testing are reported, including means, standard deviations, reliabilities, scale intercorrelations, factor analyses results, gender and race differences, and, when available, correlations with marker tests. Finally, the non-cognitive measures and the results obtained with them are summarized.

TEMPERAMENT/BIODATA CONSTRUCTS

Before discussing constructs that underlie the development of the ABLE, we need to explain how and why the inventory combines the two domains of temperament and biodata. Primarily, this action was taken to capitalize on the complementary strengths and weaknesses of each domain. The differences that exist between them allow each to contribute unique information to an assessment, and yet are not so large as to preclude a unified inventory, as described in Chapter 1.

Temperament and biodata differ from each other along the sign/sample continuum proposed by Wernimont and Campbell (1968). Biodata items are best viewed as a sample of past behavior that may predict future behavior in a similar situation. Temperament measures are most often a sign, or an indicator, of a predisposition to behave in certain ways. Thus, each type of information is geared toward predicting future behavior, but each does it from a somewhat different perspective along the sign/sample continuum.

Temperament and biodata may also differ in the emphasis placed on conceptual understanding. The study of temperament has, over the years, attached importance to the measurement of constructs and the understanding associated with such measurement. Biodata, by contrast, has typically been employed in situations requiring maximal criterion-related validity but little resulting understanding.

In short, temperament and biodata both are used to predict an individual's future behavior, but from different viewpoints and perhaps for differing reasons. The distinctions between items from the two domains are not sharp, so merging of the two sets is feasible. Yet their respective strengths complement each other when combined in a unified fashion, as in the ABLE.

In this section, we discuss the six temperament/biodata constructs assessed by the ABLE, the physical condition constructs and the response validity scales that were developed. Table 7.1 shows these eight categories and the 15 scales that fall under them.

Strictly speaking, the physical condition construct does not fit into the temperament/biodata domain in the same way that the other constructs do. It is a highly specific construct that does not have the relatively extensive, prior research history that the other constructs have. It was included, however, because the construct was seen as important for Army occupations and because we could not measure physical condition directly as part of this research project. The ABLE seemed the best instrument for collecting the physical condition measure, and so it was included as one of the target constructs.

When used in the initial pilot testing at Fort Campbell, the ABLE included a total of 291 items. It was shortened to 268 items for the later Fort Lewis pilot test. (See Chapter 2 for detailed information on the procedures and samples for these pilot tests.) Most of these items have three response options that reflect a continuum of the construct in question. The response option that reflects the highest level of the construct

Table 7.1

Temperament/Biodata Scales (by Construct) Developed for Pilot Trial Battery: ABLE - Assessment of Background and Life Experiences

Construct	Scale
Adjustment	Emotional Stability
Dependability	Nondelinquency Traditional Values Conscientiousness
Achievement	Work Orientation Self-Esteem
Physical Condition	Physical Condition
Leadership (Potency)	Dominance Energy Level
Locus of Control	Internal Control
Agreeableness/Likeability	Cooperativeness
Response Validity Scales	Non-Random Response Unlikely Virtues (Social Desirability) Poor Impression Self-Knowledge

(e.g., most dominant) is scored as a 3, while the middle response option is scored as a 2 and the lowest level response is scored as a 1. The direction of scoring differs from item to item, so the first response option is sometimes high on the construct (i.e., scored as a 3) and sometimes low (scored as a 1), to prevent response bias.

We now discuss each construct in turn and the scales developed to tap that construct. The description of the number of items on each scale refers to the Fort Campbell version.

Adjustment

Adjustment is defined as the amount of emotional stability and stress tolerance that one possesses. The well-adjusted person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He or she thinks clearly and maintains composure and rationality in situations of actual or perceived stress. The poorly adjusted person is nervous, moody, and easily irritated, tends to worry a lot, and "goes to pieces" in times of stress.

The scale included under the Adjustment construct is called Emotional Stability. It is a 31-item scale that contains items such as:

- Have you ever felt sick to your stomach when you thought about something you had to do?"
- Do you handle pressure better than most other people?

The scale is designed to assess a person's characteristic affect and ability to cope effectively with stress.

Dependability

The Dependability construct refers to a person's characteristic degree of conscientiousness. The dependable person is disciplined, well-organized, planful, respectful of laws and regulations, honest, trustworthy, wholesome, and accepting of authority. Such a person prefers order and thinks before acting. The less dependable person is unreliable, acts on the spur of the moment, and is rebellious and contemptuous of laws and regulations. Three ABLE scales fall under the Dependability construct: including Nondelinquency, Traditional Values, and Conscientiousness.

Nondelinquency is a 24-item scale that assesses how often a person has violated rules, laws, or social norms. It includes items such as:

- How often have you gotten into fights?
- Before joining the Army, how hard did you think learning to take orders would be?
- How many times were you suspended or expelled from high school?

Traditional Values, a 19-item scale under the Dependability construct, contains items such as the following:

- Are you more strict about right and wrong than most people your age?
- People should have greater respect for authority. Do you agree?

These items assess how conventional or strict a person's value system is, and how much flexibility he/she has in this value system.

Conscientiousness, the third scale falling under the Dependability construct, contains 24 items. This scale assesses the respondent's degree of dependability, as well as the tendency to be organized and planful. Items include:

- How often do you keep the promises you make?
- How often do you act on the spur of the moment?

Achievement

The Achievement construct is defined as the tendency to strive for competence in one's work. The achievement/work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in completion of the task at hand. This person is also confident, feels success from past undertakings, and expects to succeed in the future. The less achievement-oriented person has little ego involvement in his or her work, feels incapable and self-doubting, does not expend much effort, and does not feel that hard work is desirable. Two scales fall under the Achievement construct.

The 31-item scale entitled Work Orientation addresses how long, hard, and well the respondent typically works and also how he/she feels about work. Among the scale items are these:

- How often do you give up on a difficult problem?
- How hard were you willing to work for good grades in high school?
- How important is your work to you?

The other scale pertaining to Achievement is called Self-Esteem, a 16-item scale that measures how much a person believes in himself/herself and how successful he/she expects to be in life. Items from this scale include:

- Do you believe you have a lot to offer the Army?
- Has your life so far been pretty much a failure?

Physical Condition

The optimal way to establish physical condition is, of course, to administer physical conditioning tests. Since such a program could not be a part of the Trial Battery, however, it was decided to ask self-report questions through which soldiers could indicate their perceived physical fitness levels. As noted earlier, the construct of physical condition was included in the ABLE because it was the best tool available to collect such self-report data.

The Physical Condition construct refers to one's frequency and degree of participation in sports, exercise, and physical activity. Individuals high on this dimension actively participate in individual and team sports and/or exercise vigorously several times per week. Those low on this dimension have participated only minimally in athletics, exercise infrequently, and prefer the elevator to the stairs.

The scale developed to tap this construct is also called Physical Condition, and includes 14 items. The items assess how vigorously, regularly, and well the respondent engages in physical activity. These items are included on the scale:

- Prior to joining the Army, how did your physical activity (work and recreation) compare to most people your age?

- Before joining the Army, how would you have rated your performance in physical activities?

Leadership (Potency)

This construct is defined as the degree of impact, influence, and energy that one displays in relation to other people. The person high on this characteristic is appropriately forceful and persuasive, is optimistic and vital, and has the energy to get things done. The person low on this characteristic is timid about offering opinions or providing direction and is likely to be lethargic and pessimistic.

Two ABLE scales are associated with the Leadership construct: Dominance and Energy Level. Dominance is a 17-item scale that includes such items as:

- How confident are you when you tell others what to do?
- How often do people turn to you when decisions have to be made?

The scale assesses the respondent's tendency to take charge or to assume a central and public role.

The other Leadership scale, entitled Energy Level, is designed to measure to what degree one is energetic, alert, and enthusiastic. This scale includes 27 items, such as these:

- Do you get tired pretty easily?
- At what speed do you like to work?
- Do you enjoy just about everything you do?

Locus of Control

The Locus of Control construct refers to one's characteristic belief in the amount of control he/she has or people have over rewards and punishments. The person with an internal locus of control expects that there are consequences associated with behavior and that people control what happens to them by what they do. The person with an external locus of control believes that what happens to people is beyond their personal control.

The Internal Control scale is the only ABLE scale that taps the Locus of Control construct. It is a 21-item scale that assesses both internal and external control, primarily as they pertain to reaching success on the job and in life. The following are example items:

- Getting a raise or a promotion is usually a matter of luck. Do you agree?
- Do you believe you can get most of the things you want if you work hard enough for them?

Agreeableness/Likeability

The Agreeableness/Likeability construct is defined as the degree of pleasantness versus unpleasantness a person exhibits in interpersonal relations. The agreeable and likeable person is pleasant, tolerant, tactful, helpful, not defensive, and generally easy to get along with. His/her participation in a group adds cohesiveness rather than friction. The relatively disagreeable and unlikeable person is critical, fault-finding, touchy, defensive, alienated, and generally contrary.

The Cooperativeness scale is the only measure of this construct in the ABLE, and is composed of 28 items. These items assess how easy it is to get along with the person making the responses. Items from this scale include:

- How often do you lose your temper?
- Would most people describe you as pleasant?
- How well do you accept criticism?

Response Validity Scales

The purpose of the validity scales is to provide additional information about the way in which respondents have completed the ABLE. The primary purpose of these scales is to determine the validity of the responses, that is, the degree to which the responses are accurate depictions of the person completing the inventory. Those who are responding in an inaccurate way can be identified, and appropriate action taken. (For example, scores on content scales could be adjusted or the subject could be required to retake the inventory.) For those who appear to be responding accurately, the responses can be analyzed with greater confidence.

Four validity scales are included on the ABLE: Non-Random Response, Unlikely Virtues (Social Desirability), Poor Impression, and Self-Knowledge. These validity scales are modeled on similar kinds of scales that are routinely used in many measures of temperament, for example on the Minnesota Multiphasic Psychological Inventory (Dahlstrom, Welsh, & Dahlstrom, 1975) and the California Psychological Inventory (Gough, 1975). Each scale is discussed below.

The Non-Random Response scale is very different in content and scoring from other scales in the ABLE. The response options for an item do not form a continuum that indicates more or less random responding. Rather, there is one right answer which is scored as a 1, while the other two response options are both wrong and are both scored zero. Also, the content does not ask about oneself; instead, it asks about information that any person is virtually certain to know.

Two of the eight items from the Non-Random Response scale are shown next:

- The branch of the military that deals most with airplanes is the:

  1. Military Police
  2. Coast Guard
  3. Air Force

- Groups of soldiers are called:

  1. Tribes
  2. Troops
  3. Weapons

The intent of this scale is to detect those respondents who cannot or are not reading the questions, and are instead randomly filling in the circles on the answer sheet. Responses from those with a low score on this scale may be eliminated from the analyses since their responses appear to be random.

The second validity scale, entitled Unlikely Virtues, is aimed at detecting those who respond in a socially desirable manner (i.e., "fake good") rather than an honest manner. There are 12 items on this scale, of which these are a sample:

- Do you sometimes wish you had more money?

- Have you always helped people without even the slightest bit of hesitation?

Scoring on this scale uses the continuum of response options as described earlier, and those with a high score appear to be responding as they think a person should rather than honestly.

Poor Impression is the third of the ABLE validity scales, and reflects attempts to simulate psychopathology. Persons who attempt to "fake bad" receive the most deviant scores on scales such as this, while psychiatric patients score average or slightly higher than average. Thus, this scale is designed to detect those respondents who wish to make themselves appear emotionally unstable when in fact they are not unstable.

The Poor Impression scale has 23 items, most of which are also scored on another substantive ABLE scale. Items from this scale include the following:

- How much resentment do you feel when you don't get your way?

- Did your high school classmates consider you easy to get along with?

- How often do you keep the promises that you make?

Scoring on the scale is similar to that of the Non-Random Response scale, in which only one of the response options is scored as a 1 and the other two response options are scored zero. The response option scored 1 is the option that indicates the least social desirability.

278

The final validity scale is the Self-Knowledge scale, which has 13 items. This scale is intended to identify people who are more self-aware, more insightful, and more likely to have accurate perceptions about themselves. The responses of persons high on this scale may have more validity for predicting job criteria. The following are items from the Self-Knowledge scale:

- Do other people know you better than you know yourself?
- How often do you think about who you are?

All three of these scales (Unlikely Virtures, Poor Impression, and Self-Knowledge) could be used to identify suspect inventories in order to either drop the inventory from further analysis or adjust the content scales to take account of the scores on these scales. It was part of the research task to collect and analyze data to inform the best way to use these scales. In particular, the faking/fakability research, reported in Chapter 8, was intended to fulfill this purpose.

ABLE REVISIONS BASED ON PILOT TESTING

The non-cognitive inventories were pilot tested at two of the three pilot test sites: Fort Campbell and Fort Lewis. Data from these pilot tests are presented in the following section. First, however, in the following paragraphs we discuss the changes made in the ABLE on the basis of the two pilot tests to prepare the ABLE inventory for field testing. The changes are discussed for the ABLE as a whole rather than by scale, since the changes made were highly similar across scales.

Revision of the ABLE took place in three steps. The first was editorial revision prior to pilot testing, the second was based on Fort Campbell results, and the third was based on Fort Lewis findings. The editorial changes prior to pilot testing were made by PDRI, acting on suggestions from both ARI and PDRI reviews of the instrument.

The first editorial review resulted in the deletion of 17 items and the revision of 158 items. These actions were made to improve the apparent quality of the inventory, and largely consisted of minor changes in wording. Many of the changes resulted in more consistency across items in format, phrasing, and response options, and made the inventory easier and faster to take.

When the inventory was initially administered at Fort Campbell on 16 May 1984, the respondents raised very few criticisms or concerns about the ABLE. Several subjects did note the redundancy of the items on the Physical Condition scale, and this 14-item scale was shortened to nine items. One additional item characterized as irrelevant was revised.

Item analyses were based on data from 52 Fort Campbell subjects who completed the ABLE. The two statistics that were examined for each ABLE item were its correlation with the total scale on which it is scored and the endorsement frequencies for all of its response options.

Items that failed to correlate at least .15 in the appropriate direction with their respective scales were considered potentially weak. Items, other than validity scale items, for which one or more of the response options were endorsed by fewer than two subjects (i.e., < 4% of the sample) were also identified. Six items fell into the former category, 63 items fell into the latter, and an additional 7 fell into both. All of them were examined for revision or deletion, as appropriate.

In summary, a total of 23 items were deleted and 173 items revised on the basis of the editorial review and Fort Campbell findings. Items deleted were those that did not "fit well" either conceptually or statistically, or both, with the other items in the scale and with the construct in question. If the item appeared to have a "good fit" but was not clear or did not elicit sufficient variance, it was revised rather than deleted. The ABLE, which had begun at 291 items, was now a revised 268-item inventory ready to be administered at Fort Lewis.

The ABLE inventory was completed by 118 soldiers during the 11-15 June pilot testing at Fort Lewis. Item response frequency distributions were examined to detect items with relatively little discriminatory power. There were only three items where two of the three response choices were endorsed by less than 10% of the sample (not including validity scale items). After examining the content of these three items, it was decided to leave two of them intact, and delete one. Twenty items were revised because one of the three response choices was endorsed by less than 10 percent of the sample.

Overall, the inventory appeared to be functioning well and only minor revisions were required prior to field test. On the following pages, the psychometric data obtained during the two pilot tests are presented.

PILOT TEST DATA FOR THE ABLE

Fort Campbell

We begin the presentation of Fort Campbell pilot test data with the results of data screening for the ABLE. The responses of four soldiers were eliminated from analyses--two because more than 10 percent of the data was missing, and two because their Non-Random Response scale scores suggested possible random responding (scores less than 7, out of 8). The total N remaining was .52.

Table 7.2 presents means, standard deviations, mean item-total correlations, and Hoyt internal consistency reliabilities for each ABLE scale. The Poor Impression scale is not shown in this table because it was not scored for this sample. This scale is made up almost entirely from items appearing on other scales and, as described earlier, was intended to detect respondents trying to simulate psychopathology--usually for purposes of avoiding entry into the military. Since these subjects were volunteers currently on active duty, the sample size was small, and we had invoked no experimental conditions designed to elicit a range of scores on this scale. We, therefore, did not score or analyze this scale on this sample.

The reliabilities of the ABLE scales are excellent. In Table 7.3, the scale intercorrelations are shown. It is interesting to note the low correlations between the Unlikely Virtues scale, which is an indicator of Social Desirability, and the other scales. This finding, although based on a small sample, suggests that soldiers were not responding only in a socially desirable fashion, but instead were responding honestly.

The matrix of 10 ABLE scale intercorrelations (Physical Condition and the validity scales were not included) was factor analyzed (principal factor analysis) and rotated to a simple structure (varimax rotation). The four-factor solution that appeared most meaningful is shown in Table 7.4. We labeled the four factors Potency, Socialization, Dependability, and Likeability.

The scales loading highest on Factor I, Potency, are Dominance, Energy Level, and Self-Esteem; the scales loading highest on Factor II, Socialization, are Locus of Control, Traditional Values, and Nondelinquency; the scales loading highest on Factor III, Dependability, are Conscientiousness and Work Orientation; the scales loading highest on Factor IV, Likeability, are Emotional Stability and Cooperativeness. These results are, however, viewed as extremely tentative, given the small sample size upon which the factor analysis was based.

In addition to the ABLE, four well-established measures of temperament had been administered to 46 Fort Campbell soldiers to serve as marker variables: the Socialization scale of the California Psychological Inventory, Rotter's Locus of Control scale, and the Stress Reaction scale and Social Potency scale of the Differential Personality Questionnaire. The four scales (known as the Personal Opinion Inventory, POI) had also been used earlier in this project as part of the Preliminary Battery.

Data screening for this joint administration of the ABLE and the POI marker variables results in elimination of three inventories (two on the ABLE and one on the POI) because more than 10 percent of the data was missing, and

Table 7.2

Fort Campbell Pilot Test: ABLE Scale Statistics
(N = 52)

	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
ABLE Substantive Scale					
ADJUSTMENT					
Emotional Stability	31	72.06	9.10	.47	.87
DEPENDABILITY					
Nondelinquency	24	55.90	6.28	.40	.80
Traditional Values	19	43.77	4.81	.39	.73
Conscientiousness	24	58.04	5.83	.41	.80
ACHIEVEMENT					
Work Orientation	31	74.46	8.02	.42	.84
Self-Esteem	16	37.35	5.03	.54	.84
LEADERSHIP (POTENCY)					
Dominance	17	37.67	5.04	.53	.78
Energy Level	27	61.29	7.19	.46	.85
LOCUS OF CONTROL					
Internal Control	21	50.98	6.34	.46	.84
AGREEABLENESS/LIKEABILITY					
Cooperativeness	28	63.81	6.99	.39	.82
PHYSICAL CONDITION					
Physical Condition	14	43.08	9.66	.66	.92
ABLE Response Validity Scale					
Non-Random Response	8	--	--	--	--
Unlikely Virtues	12	17.98	3.19	.38	.37
Self-Knowledge	13	31.42	3.68	.43	.61

Table 7.3

Fort Campbell Pilot Test: ABLE Scale Intercorrelations
(N = 52)

	Emotional Stability	Nondelinquency	Traditional Values	Conscientiousness	Work Orientation	Self-Esteem	Dominance	Energy Level	Internal Control	Cooperativeness	Physical Condition	Unlikely Virtues	Self-Knowledge
Emotional Stability	--	45	51	42	42	61	42	53	47	56	22	06	13
Nondelinquency	45	--	71	67	51	53	25	33	58	52	01	13	31
Traditional Values	51	71	--	58	59	56	33	54	70	56	23	19	24
Conscientiousness	42	67	58	--	79	68	44	61	53	53	20	09	40
Work Orientation	42	51	59	79	--	72	52	77	59	47	18	10	39
Self-Esteem	61	53	56	68	72	--	65	73	62	41	26	10	22
Dominance	42	25	33	44	52	65	--	62	34	08	35	-03	23
Energy Level	53	33	54	61	77	73	62	--	55	38	27	15	21
Internal Control	47	58	70	53	59	62	34	55	--	42	06	-03	27
Cooperativeness	56	52	56	53	47	41	08	38	42	--	11	16	14
Physical Condition	22	01	23	20	18	26	35	27	06	11	--	06	02
Unlikely Virtues	06	13	19	09	10	10	-03	15	-03	16	06	--	-09
Self-Knowledge	13	31	24	40	39	22	23	21	27	14	02	-09	--

NOTE: Decimals have been omitted.

Table 7.4

Fort Campbell Pilot Test: Varimax Rotated Principal Factor Analyses of 10 ABLE Scales

ABLE Scale	Factor Potency I	 Socialization II	 Dependability III	 Likeability IV
Dominance	.89	.12	.11	.00
Energy Level	.73	.20	.42	.24
Self-Esteem	.70	.39	.33	.26
Internal Control	.35	.80	.14	.14
Traditional Values	.22	.79	.24	.29
Nondelinquency	.04	.78	.36	.22
Conscientiousness	.32	.39	.77	.17
Work Orientation	.51	.32	.71	.13
Emotional Stability	.46	.29	-.05	.78
Cooperativeness	-.07	.29	.46	.78

of five inventories (two on the ABLE and three on the POI) because of low Non-Random Response scores (less than 7, out of 8, on the ABLE, and more than 3, out of 10, on the POI). Thus, the responses of 38 were used to compute correlations between ABLE scales and the markers.

Results are shown in Table 7.5. It can be seen that a given ABLE construct or scale correlates most highly with the appropriate marker variable, that is, the marker for the construct to be measured. For example, the ABLE Dominance scale correlates much higher with DPQ Social Potency (.67) than with the other three marker scales which are not related to the Dominance construct (.24, .18, .22). While these results are based on a small sample, they do indicate that the ABLE scales appear to be measuring the constructs they were intended to measure.

Table 7.5

Fort Campbell Pilot Test: Correlations Between ABLE Constructs and Scales and Personal Opinion Inventory (POI) Marker Variables[a]
(N = 38)

	POI Scale			
ABLE Construct	DPQ Stress Reaction	DPQ Social Potency	Rotter Locus of Control	CPI Socialization
Emotional Stability	$\boxed{-.70}$	.32	.30	.32
Dominance	-.24	$\boxed{.67}$	.18	.22
Internal Control	-.32	.26	$\boxed{.67}$	.60
Nondelinquency	-.34	.10	.32	$\boxed{.62}$

[a] "Marker" correlations are indicated by a box.

Fort Lewis

Soldiers at the Fort Lewis pilot test in June 1984 completed the revised version of the ABLE along with the AVOICE, the cognitive tests, and the psychomotor tests that comprised the entire Pilot Trial Battery. The final N for statistical analyses of the ABLE was 106; 1 inventory was eliminated because more than 10 percent of the data was missing, and 11 were eliminated because Non-Random Response was less than 7 (out of 8).

The means, standard deviations, mean item-total scale correlations, and Hoyt reliability estimates appear in Table 7.6 for the entire group (after screening). (Again, Poor Impression scale scores were not computed for reasons stated earlier.) As can be seen, the reliabilities of the ABLE scales are again excellent.

Tables 7.7 and 7.8 show the scale means and standard deviations for males and females, and blacks and whites, respectively. Note that the Ns are quite small for females and blacks, but these statistics do not show any striking differences between subgroups.

In Table 7.9, the scale intercorrelations are presented for all ABLE scales except the Non-Random Response and Poor Impression validity scales. It can be seen that in the Fort Lewis data, Unlikely Virtues (Social Desirability) correlates more highly with other scales than in the Fort Campbell

Table 7.6

Fort Lewis Pilot Test: ABLE Scale Statistics for Total Group

	No. Items	N	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
ABLE Substantive Scales						
ADJUSTMENT						
Emotional Stability	30	106	68.97	8.59	.46	.87
DEPENDABILITY						
Nondelinquency	25	106	59.07	6.28	.40	.78
Traditional Values	16	106	37.39	4.25	.41	.67
Conscientiousness	21	106	50.24	5.31	.41	.75
ACHIEVEMENT						
Work Orientation	27	106	62.88	7.77	.48	.86
Self-Esteem	15	106	34.90	4.71	.52	.80
LEADERSHIP (POTENCY)						
Dominance	16	106	36.55	6.08	.57	.86
Energy Level	25	106	59.26	7.40	.52	.88
LOCUS OF CONTROL						
Internal Control	21	106	49.90	6.27	.46	.80
AGREEABLENESS/LIKEABILITY						
Cooperativeness	25	106	56.41	6.70	.43	.81
PHYSICAL CONDITION						
Physical Condition	9	106	31.30	6.96	.73	.87
ABLE Validity Scales						
Non-Random Response	8	117	7.55	.71	.43	--
Unlikely Virtues	12	106	16.63	3.45	.48	.71
Self-Knowledge	13	106	29.75	3.96	.46	.71

Table 7.7

Fort Lewis Pilot Test: ABLE Scale Means and Standard Deviations Separately for Males and Females

	Males (N = 87)		Females (N = 19)	
	Mean	SD	Mean	SD
ABLE Substantive Scales				
ADJUSTMENT				
Emotional Stability	69.78	8.88	65.26	5.82
DEPENDABILITY				
Nondelinquency	58.46	6.28	61.84	5.46
Traditional Values	37.13	4.38	38.58	3.30
Conscientiousness	49.95	5.49	51.53	4.18
ACHIEVEMENT				
Work Orientation	62.17	7.78	66.11	6.89
Self-Esteem	34.72	4.73	35.68	4.53
LEADERSHIP (POTENCY)				
Dominance	36.66	6.10	36.05	5.95
Energy Level	59.21	7.65	59.53	6.12
LOCUS OF CONTROL				
Internal Control	49.66	6.31	51.00	5.93
AGREEABLENESS/LIKEABILITY				
Cooperativeness	55.93	6.99	58.58	4.61
PHYSICAL CONDITION				
Physical Condition	31.64	6.20	29.74	9.54
ABLE Validity Scales				
Non-Random Response[a]	7.50	.72	7.76	.61
Unlikely Virtues	16.63	3.57	16.63	2.81
Self-Knowledge	29.54	4.00	30.74	3.64

[a] Scale means and standard deviations are given here for data which are unscreened with respect to this scale. Thus, the N for males is 96 and for females is 21.

Table 7.8

Fort Lewis Pilot Test: ABLE Scale Means and Standard Deviations Separately for Blacks and Whites

	Blacks (N = 26)		Whites (N = 63)	
	Mean	SD	Mean	SD
ABLE Substantive Scales				
ADJUSTMENT				
Emotional Stability	66.15	7.65	70.56	8.36
DEPENDABILITY				
Nondelinquency	60.65	6.06	58.86	6.37
Traditional Values	37.50	2.96	37.86	4.66
Conscientiousness	50.69	4.45	50.29	5.76
ACHIEVEMENT				
Work Orientation	63.50	6.40	62.73	8.63
Self-Esteem	34.54	4.25	35.29	4.88
LEADERSHIP (POTENCY)				
Dominance	37.77	3.43	36.75	6.80
Energy Level	57.35	5.84	59.83	8.26
LOCUS OF CONTROL				
Internal Control	49.69	4.74	50.35	6.81
AGREEABLENESS/LIKEABILITY				
Cooperativeness	57.81	5.42	56.08	7.13
PHYSICAL CONDITION				
Physical Condition	31.92	5.94	30.95	7.11
ABLE Validity Scales				
Non-Random Response[a]	7.40	.80	7.69	.52
Unlikely Virtues	16.15	2.74	16.63	3.68
Self-Knowledge	31.23	3.46	29.43	4.09

[a] Scale means and standard deviations are given here for data which are unscreened with respect to this scale. Thus, the N is 30 for Blacks and 65 for Whites.

Table 7.9

Fort Lewis Pilot Test: ABLE Scale Intercorrelations
(N = 106)

	Emotional Stability	Nondelinquency	Traditional Values	Conscientiousness	Work Orientation	Self-Esteem	Dominance	Energy Level	Internal Control	Cooperativeness	Physical Condition	Unlikely Virtues	Self-Knowledge
Emotional Stability	--	35	29	32	40	56	36	69	50	50	15	32	11
Nondelinquency	35	--	68	63	61	40	25	46	55	61	22	42	13
Traditional Values	29	68	--	55	54	41	31	45	64	50	22	31	10
Conscientiousness	32	63	55	--	76	59	45	60	53	40	35	46	31
Work Orientation	40	61	54	76	--	74	58	74	54	44	35	41	34
Self-Esteem	56	40	41	59	74	--	64	72	57	44	33	31	32
Dominance	36	25	31	45	58	64	--	53	38	17	37	15	25
Energy Level	69	46	45	60	74	72	53	--	62	50	30	35	28
Internal Control	50	55	64	53	54	57	38	62	--	63	13	21	28
Cooperativeness	50	61	50	40	44	44	17	50	63	--	19	30	28
Physical Condition	15	22	22	35	35	33	37	30	13	19	--	14	17
Unlikely Virtues	32	42	31	46	41	31	15	35	21	30	14	--	-03
Self-Knowledge	11	13	10	31	34	32	25	28	28	28	17	-03	--

NOTE: Decimals have been omitted.

data. Table 7.10 presents the scale intercorrelations for the ten ABLE substantive scales (excluding the validity and Physical Condition scales) with Social Desirability variance partialed out. As would be expected given the correlation between Unlikely Virtues and the other ABLE scales, the values in Table 7.10 are from 3 to 10 points lower than in Table 7.9. There is no readily apparent explanation for the differences in findings between the Fort Campbell and Fort Lewis samples except for sampling error, since both sample sizes are relatively small.

Correlation matrices for the ten ABLE substantive scales from Fort Lewis were factor analyzed, both with and without the Social Desirability variance. Principal factor analyses were used, with rotation to simple structure by varimax rotation. Both factor matrices appear in Table 7.11. Though neither structure is the same as was obtained when we factor analyzed the Fort Campbell correlation matrix, the factor solution resulting when Social Desirability is partialed out is quite similar to the solution obtained with the Fort Campbell data. The differences are that in the Fort Lewis solution, Energy Level loads on a factor with Emotional Stability, whereas in the Fort Campbell solution, Energy Level loads with Dominance and Self-Esteem. The other difference is that in the five-factor Fort Lewis solution, Cooperativeness forms a factor by itself, whereas in the four-factor Fort Campbell solution, Cooperativeness forms a factor with Emotional Stability.

The structure of the temperament and biodata domain, as measured by the ABLE during the pilot tests, could not be specified with certainty due to the relatively small pilot test sample upon which the correlational and factor analyses were run. The scales do, however, appear to be measuring the same content as the corresponding marker variables that were a part of the Preliminary Battery. The internal consistency reliabilities and score distributions of the ABLE scales are more than acceptable.

Table 7.10

Fort Lewis Pilot Test: ABLE Scale Intercorrelations With Social Desirability Variance Partialed Out

	Emotional Stability	Nondelinquency	Traditional Values	Conscientiousness	Work Orientation	Self-Esteem	Dominance	Energy Level	Internal Control	Cooperativeness
Emotional Stability	--	25	21	20	31	51	37	65	47	45
Nondelinquency	25	--	63	54	52	32	21	37	52	55
Traditional Values	21	63	--	49	48	35	28	39	61	45
Conscientiousness	20	54	49	--	70	53	44	53	50	31
Work Orientation	31	52	48	70	--	71	57	69	51	36
Self-Esteem	51	32	35	53	71	--	63	69	54	39
Dominance	37	21	28	44	57	63	--	52	36	14
Energy Level	65	37	39	53	69	69	52	--	60	44
Internal Control	47	52	61	50	51	54	36	60	--	61
Cooperativeness	45	55	45	31	36	39	14	44	61	--

NOTE: Decimals have been omitted.

Table 7.11

Fort Lewis Pilot Test: Varimax Rotated Principal Factor Analyses of 10 ABLE Scales

	Five-Factor Solution With Social Desirability Variance Included				
ABLE Scale	I	II	III	IV	V
Dominance	.66	.15	.16	.00	.21
Energy Level	.45	.19	.32	.22	.79
Self-Esteem	.80	.13	.22	.30	.27
Internal Control	.33	.52	.15	.44	.29
Traditional Values	.18	.78	.29	.22	.10
Nondelinquency	.09	.50	.56	.41	.09
Conscientiousness	.40	.34	.61	.14	.16
Work Orientation	.57	.25	.63	.15	.24
Emotional Stability	.33	.11	.02	.43	.53
Cooperativeness	.08	.30	.21	.77	.22
	Five-Factor Solution With Social Desirability Variance Partialed Out				
	I	II	III	IV	V
Dominance	.65	.15	.23	-.03	.18
Energy Level	.39	.18	.82	.13	.36
Self-Esteem	.79	.12	.32	.24	.19
Internal Control	.31	.52	.34	.40	.14
Traditional Values	.17	.83	.10	.17	.18
Nondelinquency	.08	.56	.06	.40	.42
Conscientiousness	.40	.37	.11	.11	.56
Work Orientation	.57	.27	.22	.13	.62
Emotional Stability	.30	.08	.60	.35	-.06
Cooperativeness	.06	.31	.26	.78	.13

INTERESTS CONSTRUCTS

The seminal work of John Holland (1966) has resulted in widespread acceptance of a six-construct, hexagonal model of interests. Our principal problem in developing and testing an interests measure for Army testing was not which constructs to measure, but rather how much emphasis should be devoted to the assessment of each.

As earlier stated, the interests inventory that had been used in the Preliminary Battery is called the VOICE (Vocational Interest Career Examination), which had been developed and researched by the U.S. Air Force. This inventory served as the starting point for the AVOICE (Army Vocational Interest Career Examination).

When developing the AVOICE, we sought to ensure that it would measure well all six of Holland's constructs, as well as provide sufficient coverage of the vocational areas most important in the Army. We wanted the inventory's items to parallel the job tasks of soldiers in a variety of MOS, while at the same time assessing a respondent's broad interests. Thus, each of the constructs to be discussed next is adequately measured by the AVOICE; however, a greater degree of coverage is devoted to constructs judged most important for Army jobs. Table 7.12 shows the six Holland interests constructs assessed by the AVOICE, together with their associated scales.

In addition to the Holland constructs and associated scales, the AVOICE also included six constructs (20 scales) dealing with organizational climate and environment preferences and an expressed interests scale. Table 7.13 shows these variables and associated measures.

As used in the pilot testing, the AVOICE included 306 items. Nearly all items were scored on a 5-point scale that ranged from "Like Very Much" (scored 5) to "Dislike Very Much" (scored 1). Items in the Expressed Interests scale were scored on a 3-point scale in which the response options were different for each item, yet one option always reflected the most interest, one moderate interest, and one the least interest.

We now discuss, in turn, each construct/category and the scales developed for it.

Realistic Interests

This construct is defined as a preference for concrete and tangible activities, characteristics, and tasks. Persons with realistic interests enjoy and are skilled in the manipulation of tools, machines, and animals, but find social and educational activities and situations aversive. Realistic interests are associated with occupations such as mechanic, engineer, and wildlife conservation officer, and negatively associated with such occupations as social work and artist.

The Realistic construct is by far the most thoroughly assessed of the six constructs tapped by the AVOICE, reflecting the preponderance of work in the Army of a Realistic nature. Fourteen AVOICE scales fall under this construct, in addition to a Basic Interest item.

Table 7.12

Holland Basic Interest Constructs, and Army Vocational Interest Career Examination Scales Developed for Pilot Trial Battery: AVOICE - Army Vocational Interest Career Examination

Construct	Scale
Realistic	Basic Interest Item Mechanics Heavy Construction Electronics Electronic Communication Drafting Law Enforcement Audiographics Agriculture Outdoors Marksman Infantry Armor/Cannon Vehicle Operator Adventure
Conventional	Basic Interest Item Office Administration Supply Administration Food Service
Social	Basic Interest Item Teaching/Counseling
Investigative	Basic Interest Item Medical Services Mathematics Science/Chemical Automated Data Processing
Enterprising	Basic Interest Item Leadership
Artistic	Basic Interest Item Aesthetics

Table 7.13

Additional AVOICE Measures: Organizational Climate/Environment and Expressed Interests Scales

Construct	Scale
Achievement (Org. Climate/Environment)	Achievement Authority Ability Utilization
Safety (Org. Climate/Environment)	Organizational Policies and Procedures Supervision - Human Resources Supervision - Technical
Comfort (Org. Climate/Environment)	Activity Variety Compensation Security Working Conditions
Status (Org. Climate/Environment)	Advancement Recognition Social Status
Altruism (Org. Climate/Environment)	Co-workers Moral Values Social Services
Autonomy (Org. Climate/Environment)	Responsibility Creativity Independence
Expressed Interests	Expressed Interests

The Basic Interest item, one of which is written for each Holland construct, describes a person with prototypic Realistic interests. The respondent indicates how well this description fits him/her. The remaining Realistic scales are discussed next.

- The Mechanics scale is a 16-item scale that measures interest in various kinds of mechanical work. Sample items include:
  - Replace valves in an engine.
  - Adjust a carburetor.
- Heavy Construction is a 23-item scale dealing with interest in construction tasks. Example items are:
  - Mason.
  - Welder.
  - Construct a quick shelter in the woods.
- Twenty items are included on the Electronics scale. Items from this scale include these:
  - Repair a television set.
  - Design a circuit board.
  - Wiring diagrams.
- The Electronic Communication scale concerns interest in transmitting information electronically. This 7-item scale includes such items as:
  - Operate radio and teletype equipment.
  - Telecommunications.
- Drafting is also a Realistic scale with seven items. Among the Drafting scale items are:
  - Artist.
  - Draftsman.
  - Draw blueprints for a bridge.
- Another Realistic scale is called Law Enforcement and includes both security and law enforcement components. Three of the scale's 16 items are:
  - Highway patrol officer.
  - Prison guard.
  - Be a witness at a criminal trial.
- The Audiographics scale, which has seven items, concerns activities associated with photography and movies. Items from this scale are:
  - Photographer.
  - Record the sound for a motion picture.

298

- One of the shortest Realistic scales, Agriculture, contains only five items. Two of the scale's items are:
    - Drive a tractor on a farm.
    - Mow lawns, clip hedges, and trim trees.
- The Outdoors scale contains nine items including:
    - Work outdoors.
    - Go deer hunting.
    - Learn survival techniques for living in the wilderness.
- The Marksman scale's five items include:
    - Gunsmith.
    - Teach marksmanship.
    - Collect rifles and pistols.
- The Infantry scale contains ten activities engaged in by infantrymen. Among these items are:
    - Use cover, concealment, and camouflage.
    - Clear a mine field.
    - Direct artillery fire.
- Armor/Cannon is an 8-item scale that pertains to operating large ground-based weapons. The items include:
    - Zero in a tank's main gun.
    - Load and unload field artillery cannons.
- The scale entitled Vehicle Operator includes the following among its nine items:
    - Taxi driver.
    - Deliver cargo on time.
    - Operate a bulldozer or power shovel.
- Finally, the Adventure scale has eight items that include:
    - Explore a wilderness area alone.
    - Go skydiving.
    - Hunt wild animals in Africa.

Eight ABLE items are also scored on the Adventure scale. Thus, we could obtain Adventure scores based on AVOICE items only, ABLE items only, or both. In this section, we will deal only with the eight AVOICE Adventure items.

Conventional Interests

The construct of Conventional interests refers to one's degree of preference for well-ordered, systematic and practical activities and tasks. Persons with Conventional interests may be characterized as conforming, unimaginative, efficient, and calm. Conventional interests are associated with occupations such as accountant, clerk, and statistician, and negatively associated with occupations such as artist or author.

In addition to the Basic Interest item, three scales fall under the Conventional interests construct--Office Administration, Supply Administration, and Food Service. They have, respectively, 16, 13, and 17 items. Example items from these three scales are:

- Office Administration -
  - Make copies of a letter.
  - Keep accurate records.
  - Schedule appointments for other people.
- Supply Administration -
  - Prepare materials, equipment, or supplies for shipment.
  - Make out invoices.
  - Take inventory for a department store.
- Food Service -
  - Dishwasher.
  - Buy food supplies for a restaurant.
  - Wash, peel and dice vegetables.

Social Interests

Social interests are defined as the amount of liking one has for social, helping, and teaching activities and tasks. Persons with social interests may be characterized as responsible, idealistic, and humanistic. Social interests are associated with occupations such as social worker, high school teacher, and speech therapist, and negatively associated with occupations such as mechanic or carpenter.

- Besides the Basic Interest item, only one scale is included in the AVOICE for assessing Social interests, the Teaching/Counseling scale. This 7-item scale includes items such as:
  - Give on-the-job training.
  - Organize and lead a study group.
  - Listen to people's problems and try to help them.

Investigative Interests

This construct refers to one's preference for scholarly, intellectual, and scientific activities and tasks. Persons with Investigative interests enjoy analytical, ambiguous, and independent tasks, but dislike leadership and persuasive activities. Investigative interests are associated with such occupations as astronomer, biologist, and mathematician, and negatively associated with occupations such as salesman or politician.

Along with the Basic Interest item, Medical Services, Mathematics, Science/Chemical, and Automated Data Processing are the four AVOICE scales that tap Investigative Interests. The scales differ in length with Medical Services containing 24 items; Mathematics, 5; Science/Chemical, 11; and, Automated Data Processing, 7. Again, selected scale items are supplied below.

- Medical Services -
  - Physical Therapist.
  - Take blood pressure readings.
  - Disease prevention.
- Mathematics -
  - Solve arithemetic problems.
  - Find information in numerical tables.
  - Work with numbers.
- Science/Chemical -
  - Mix chemical compounds.
  - Record observations from scientific instruments.
  - Work with hazardous chemicals.
- Automated Data Processing -
  - Computer Operator.
  - Computer Programmer.
  - Operate a machine that sorts punched cards.

Enterprising Interests

The Enterprising interests construct refers to one's preference for persuasive, assertive, and leadership activities and tasks. Persons with Enterprising interests may be characterized as ambitious, dominant, sociable, and self-confident. Enterprising interests are associated with such occupations as salesperson and business executive, and negatively associated with occupations such as biologist or chemist.

- Again, besides the Basic Interest item, only one AVOICE scale assesses the respondent's Enterprising interests. This scale, entitled Leadership, contains six items including the following:
  - Mold a group of coworkers into an efficient team.
  - Inspire others with a speech.
  - Make decisions when others do not know what to do.

Artistic Interests

This final Holland construct is defined as a person's degree of liking for unstructured, expressive, and ambiguous activities and tasks. Persons with Artistic interests may be characterized as intuitive, impulsive, creative, and non-conforming. Artistic interests are associated with such occupations as writer, artist, and composer, and negatively associated with occupations such as accountant or secretary.

- In addition to the Basic Interest item, the AVOICE Aesthetics scale is designed to tap Artistic Interests, and includes five items. Among these items are:

    - Read poetry.
    - Watch educational television.
    - Classical music.

Organizational Climate/Environment Scales

Six constructs that pertain to a person's preference for certain types of work environments and conditions are assessed by the AVOICE through 20-item scales. These environmental constructs include Achievement, Safety, Comfort, Status, Altruism, and Autonomy. The items that assess these constructs are distributed throughout the AVOICE, and are responded to in the same manner as the interests items, that is, "Like Very Much" to "Dislike Very Much."

Because the scales contain only two items each and for ease of presentation, Figure 7.1 is used to show the constructs, scales, and an item from each scale.

Expressed Interests Scale

Although not a psychological construct, expressed interests were included in the AVOICE because of the extensive research showing their validity in criterion-related studies. (Dolliver, 1969) These studies had measured expressed interests simply by asking respondents what occupation or occupational area was of most interest to them. In the AVOICE, such an open-ended question was not feasible, instead, respondents were asked how confident they were that their chosen job in the Army was the right one for them.

This Expressed Interests scale contained eight items which, as mentioned, had three response options that formed a continuum of confidence in the person's occupational choice. Selected items from this scale include:

- Before you went to the recruiter, how certain were you of the job you wanted in the Army?

- If you had the opportunity right now to change your job in the Army, would you?

- Before enlisting, how long were you interested in a particular Army job?

Construct/Scale	Example
Achievement	
Achievement	"Do work that gives a feeling of accomplishment."
Authority	"Tell others what to do on the job."
Ability Utilization	"Make full use of your abilities."
Safety	
Organizational Policy	"A job in which the rules are not equal for everyone."
Supervision - Human Resources	"Have a boss that supports the workers."
Supervision - Technical	"Learn the job on your own."
Comfort	
Activity	"Work on a job that keeps a person busy."
Variety	"Do something different most days at work."
Compensation	"Earn less than others do."
Security	"A job with steady employment."
Working Conditions	"Have a pleasant place to work."
Status	
Advancement	"Be able to be promoted quickly."
Recognition	"Receive awards or compliments on the job."
Social Status	"A job that does not stand out from others."
Altruism	
Co-workers	"A job in which other employees were hard to get to to know."
Moral Values	"Have a job that would not bother a person's conscience."
Social Services	"Serve others through your work."
Autonomy	
Responsibility	"Have work decisions made by others."
Creativity	"Try out your own ideas on the job."
Independence	"Work alone."

Figure 7.1. Organizational climate/environment preference constructs, scales within constructs, and an item from each scale.

AVOICE REVISIONS BASED ON PILOT TESTING

As with the ABLE, before we present the data obtained from pilot testing, we describe the revisions made in the AVOICE on the basis of pilot test administration at Fort Campbell and Fort Lewis. Again, the changes are discussed for the AVOICE as a whole, rather than scale by scale. These changes resulted in the AVOICE version to be used in the field test.

Overall, the revisions made were far less substantial for the AVOICE than for the ABLE. Editorial review of the inventory by PDRI and ARI staff, together with the verbal feedback from Fort Campbell soldiers, resulted in revision of 15 items--primarily minor wording changes. An additional five items were modified because of low item correlations with the total scale score in the Fort Campbell data. No items were deleted based on the editorial review, verbal feedback, or item analyses.

Following the Fort Lewis pilot test, no revisions or deletions were made to the AVOICE items. Item response frequencies were examined to detect items that had relatively little discriminatory power, that is, three or more of the five response choices received less than 10 percent endorsement. There proved to be only two such items, and, upon examination of the item content, it was decided not to revise these. Both items appeared well written and relevant to the targeted content, and we thought the poor response distribution could be attributed to sampling error.

Thus, a total of only 20 AVOICE items were revised on the basis of editorial review and pilot testing. Part of this low level of revision may be due to the common response scale of the inventory, "Like Very Much" to "Dislike Very Much." The response options appeared to be well-understood and did not require the item-by-item review/revision that was necessary for the ABLE items (which had differing response options by item).

PILOT TEST DATA FOR THE AVOICE

Fort Campbell

In the Fort Campbell pilot test, a total of 57 soldiers completed the AVOICE, 55 of whom provided sufficient data for analyses. Scale statistics for this sample are presented in Table 7.14. As can be seen in the table, the mean item-total correlations and Hoyt reliabilities are excellent, generally in the .60s to .80s for the former, and .70s to .90s for the latter. In addition the means and SDs indicate acceptable scale score distributions in almost all cases.

Fort Lewis

The responses of four of 118 soldiers were eliminated for exceeding the missing data criterion (10%), resulting in an analysis sample size of 114. Scale statistics for this sample are shown in Table 7.15. Reliabilities are again excellent and are even slightly higher than the values obtained at Fort Campbell.

AVOICE scale means and standard deviations were also calculated separately for males and females and for blacks and whites (see Tables 7.16 and 7.17), but note that sample sizes are very small for females and blacks. These data are viewed as exploratory only. As would be expected on the basis of previous research, there are marked differences between the sexes in mean score on certain interest scales. Scales such as Mechanics and Heavy Construction show far greater scores for males than females. On the majority of the scales, however, the differences are less pronounced. Differences are also relatively small between blacks and whites. Table 7.18 presents the AVOICE scale intercorrelations for the Fort Lewis sample. We performed no detailed analyses of these correlations, but did inspect the matrix to see if scales expected to correlate fairly highly did so (for example, Infantry with Armor/Cannon) and scales not expected to correlate highly, or even negatively, did so (for example, Aesthetics with Infantry). This pattern did indeed hold true, in most cases.

Table 7.14

Fort Campbell Pilot Test: AVOICE Scale Statistics (N = 55)

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
REALISTIC					
Basic Interest Item	1	1.95	.75	--	--
Mechanics	16	49.91	14.54	.75	.95
Heavy Construction	23	65.84	16.13	.64	.93
Electronics	20	65.45	17.48	.75	.96
Electronic Communication	7	20.00	5.15	.64	.76
Drafting	7	20.84	5.04	.62	.75
Law Enforcement	16	47.78	10.59	.55	.83
Audiographics	7	23.05	4.32	.58	.69
Agriculture	5	14.29	3.51	.60	.55
Outdoors	9	32.20	6.77	.63	.81
Marksman	5	15.25	4.64	.77	.82
Infantry	10	26.93	6.66	.57	.78
Armor/Cannon	8	22.29	6.51	.71	.87
Vehicle Operator	9	24.93	7.03	.69	.87
Adventure	8	18.87	2.11	.39	--
CONVENTIONAL					
Basic Interest Item	1	2.02	.65	--	--
Office Administration	16	41.84	13.37	.74	.94
Supply Administration	13	32.64	9.88	.72	.92
Food Service	17	39.18	8.18	.49	.81
SOCIAL					
Basic Interest Item	1	2.22	.78	--	--
Teaching/Counseling	7	22.33	5.41	.67	.80
INVESTIGATIVE					
Basic Interest Item	1	1.38	.52	--	--
Medical Services	24	66.02	17.46	.66	.95
Mathematics	5	14.09	3.79	.69	.73
Science/Chemical	11	29.15	7.60	.61	.84
Automated Data Processing	7	23.69	6.12	.73	.86
ENTERPRISING					
Basic Interest Item	1	1.84	.68	--	--
Leadership	6	19.93	4.88	.69	.78
ARTISTIC					
Basic Interest Item	1	1.62	.67	--	--
Aesthetics	5	13.33	4.00	.74	.79

(Continued)

Table 7.14 (Continued)

Fort Campbell Pilot Test: AVOICE Scale Statistics

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
ACHIEVEMENT (Org. Climate/Environment)					
Achievement	2	1.76	1.60	.75	--
Authority	2	.25	1.72	.70	--
Ability Utilization	2	1.49	1.41	.76	--
SAFETY (Org. Climate/Environment)					
Organizational Policies and Procedures	2	2.09	1.27	.69	--
Supervision-Human Resources	2	2.20	1.64	.74	--
Supervision-Technical	2	.40	1.84	.68	--
COMFORT (Org. Climate/Environment)					
Activity	2	1.45	1.55	.71	--
Variety	2	1.31	1.58	.81	--
Compensation	2	2.58	1.51	.75	--
Security	2	2.85	1.30	.77	--
Working Conditions	2	1.98	1.51	.78	--
STATUS (Org. Climate/Environment)					
Advancement	2	1.67	1.45	.69	--
Recognition	2	1.20	1.81	.73	--
Social Status	2	1.42	1.69	.75	--
ALTRUISM (Org. Climate/Environment)					
Co-workers	2	2.16	1.45	.83	--
Moral Values	2	1.60	1.66	.71	--
Social Services	2	6.98	1.80	.82	--
AUTONOMY (Org. Climate/Environment)					
Responsibility	2	1.65	1.36	.66	--
Creativity	2	.91	1.38	.58	--
Independence	2	-.44	1.25	.69	--
EXPRESSED INTEREST	8	15.15	3.89	.54	.30

Table 7.15

Fort Lewis Pilot Test: AVOICE Scale Statistics for Total Group
(N = 114)

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
REALISTIC					
Basic Interest Item	1	3.09	1.17	--	--
Mechanics	16	53.02	13.13	.73	.94
Heavy Construction	23	72.57	15.64	.62	.92
Electronics	20	63.94	16.86	.75	.96
Electronic Communication	7	21.44	5.73	.73	.85
Drafting	7	22.62	6.11	.76	.87
Law Enforcement	16	50.82	11.33	.63	.89
Audiographics	7	24.30	5.12	.69	.81
Agriculture	5	15.24	3.62	.61	.58
Outdoors	9	33.09	6.25	.62	.80
Marksman	5	16.57	4.48	.79	.84
Infantry	10	31.04	7.26	.64	.84
Armor/Cannon	8	23.46	6.15	.67	.83
Vehicle Operator	10	30.45	7.10	.65	.84
Adventure	8	18.84	3.60	.57	.72
CONVENTIONAL					
Basic Interest Item	1	3.00	.92	--	--
Office Administration	16	45.39	12.61	.72	.94
Supply Administration	13	36.97	9.65	.71	.92
Food Service	17	43.46	10.53	.59	.89
SOCIAL					
Basic Interest Item	1	3.25	1.03	--	--
Teaching/Counseling	7	23.61	5.20	.71	.83
INVESTIGATIVE					
Basic Interest Item	1	3.09	.95	--	--
Medical Services	24	71.32	16.65	.66	.94
Mathematics	5	15.82	4.20	.75	.80
Science/Chemical	11	30.29	8.41	.68	.88
Automated Data Processing	7	24.29	5.78	.74	.86
ENTERPRISING					
Basic Interest Item	1	3.11	1.13	--	--
Leadership	6	20.71	4.41	.72	.81

(Continued)

Table 7.15 (Continued)

Fort Lewis Pilot Test: AVOICE Scale Statistics for Total Group
(N = 114)

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
ARTISTIC					
Basic Interest Item	1	2.99	1.27	--	--
Aesthetics	5	14.73	4.12	.74	.79
ORGANIZATIONAL CLIMATE/ ENVIRONMENT DIMENSIONS					
Achievement	6	21.09	2.95	--	--
Safety	6	21.64	3.20	--	--
Comfort	10	38.50	3.83	--	--
Status	6	21.37	2.97	--	--
Altruism	6	21.67	3.28	--	--
Autonomy	6	20.46	2.33	--	--
EXPRESSED INTEREST	8	15.71	3.19	.59	.66

Table 7.16

Fort Lewis Pilot Test: AVOICE Means and Standard Deviations Separately for Males and Females

	Males (N = 87)		Females (N = 19)	
AVOICE Scale	Mean	SD	Mean	SD
REALISTIC				
Basic Interest Item	3.24	1.13	2.35	1.11
Mechanics	54.93	12.51	44.05	12.28
Heavy Construction	75.31	13.24	59.70	19.22
Electronics	66.38	15.95	52.45	16.23
Electronic Communication	21.48	5.73	21.25	5.72
Drafting	22.97	6.11	21.00	5.83
Law Enforcement	51.72	11.41	46.60	9.95
Audiographics	24.27	5.03	24.45	5.52
Agriculture	15.46	3.59	14.20	3.57
Outdoors	33.94	5.75	29.10	6.92
Marksman	17.35	4.05	12.90	4.56
Infantry	31.94	7.14	26.85	6.28
Armor/Cannon	24.21	5.99	19.95	5.71
Vehicle Operator	31.05	6.52	27.60	8.81
Adventure	19.39	3.28	16.32	3.91
CONVENTIONAL				
Basic Interest Item	2.97	.92	3.15	.91
Office Administration	44.91	11.93	47.60	15.19
Supply Administration	36.95	9.56	37.10	10.09
Food Service	42.54	9.89	47.80	12.23
SOCIAL				
Basic Interest Item	3.24	1.05	3.30	.95
Teaching/Counseling	23.15	5.13	25.75	4.97
INVESTIGATIVE				
Basic Interest Item	3.10	.95	3.05	.97
Medical Services	71.10	16.65	72.40	16.59
Mathematics	15.59	4.31	16.95	3.40
Science/Chemical	30.99	8.69	27.00	5.96
Automated Data Processing	24.20	5.97	24.70	4.76

(Continued)

Table 7.16 (Continued)

Fort Lewis Pilot Test: AVOICE Means and Standard Deviations Separately for Males and Females

	Males (N = 87)		Females (N = 19)	
AVOICE Scale	Mean	SD	Mean	SD
ENTERPRISING				
Basic Interest Item	3.14	1.14	2.95	1.02
Leadership	20.53	4.61	21.55	3.17
ARTISTIC				
Basic Interest Item	2.96	1.25	3.15	1.31
Aesthetics	14.29	4.22	16.80	2.77
ORGANIZATIONAL CLIMATE/ ENVIRONMENT DIMENSIONS				
Achievement	20.97	2.92	21.65	3.02
Safety	21.59	3.36	21.90	2.23
Comfort	38.26	3.76	39.65	3.97
Status	21.22	3.00	22.05	2.73
Altruism	21.48	3.26	22.55	3.26
Autonomy	20.45	2.22	20.55	2.78
EXPRESSED INTEREST	15.79	3.34	15.35	2.29

Table 7.17

Fort Lewis Pilot Test: AVOICE Means and Standard Deviations Separately for Blacks and Whites

AVOICE Scale	Blacks (N = 27) Mean	Blacks (N = 27) SD	Whites (N = 65) Mean	Whites (N = 65) SD
REALISTIC				
Basic Interest Item	2.81	1.39	3.26	1.06
Mechanics	50.96	12.29	54.20	12.90
Heavy Construction	67.85	14.10	75.69	14.55
Electronics	66.33	14.94	64.20	16.77
Electronic Communication	23.22	4.37	21.38	5.82
Drafting	23.81	5.00	22.46	6.57
Law Enforcement	48.04	12.22	53.43	10.40
Audiographics	25.00	4.58	24.82	5.05
Agriculture	14.04	3.49	16.18	3.56
Outdoors	29.81	5.12	35.28	5.19
Marksman	15.48	3.47	17.54	4.51
Infantry	29.37	6.38	32.68	7.41
Armor/Cannon	22.26	5.20	24.43	6.43
Vehicle Operator	29.37	7.42	31.42	6.92
Adventure	15.58	3.32	20.11	2.70
CONVENTIONAL				
Basic Interest Item	3.07	.77	2.92	.98
Office Administration	51.37	10.00	43.65	13.45
Supply Administration	41.19	8.68	35.72	10.42
Food Service	48.74	8.52	41.63	11.04
SOCIAL				
Basic Interest Item	3.22	.92	3.28	1.07
Teaching/Counseling	25.04	4.61	23.48	5.50
INVESTIGATIVE				
Basic Interest Item	3.11	1.10	3.14	.91
Medical Services	77.81	12.88	69.35	17.68
Mathematics	17.22	4.05	15.22	4.25
Science/Chemical	29.96	6.58	31.23	9.15
Automated Data Processing	27.93	3.87	23.63	5.90
ENTERPRISING				
Basic Interest Item	3.30	1.01	3.05	1.14
Leadership	21.44	3.82	20.97	4.59

(Continued)

Table 7.17 (Continued)

Fort Lewis Pilot Test: AVOICE Means and Standard Deviations Separately for Blacks and Whites

AVOICE Scale	Blacks (N = 27) Mean	Blacks (N = 27) SD	Whites (N = 65) Mean	Whites (N = 65) SD
ARTISTIC				
Basic Interest Item	3.44	1.37	2.88	1.23
Aesthetics	15.59	3.29	14.66	4.50
ORGANIZATIONAL CLIMATE/ ENVIRONMENT DIMENSIONS				
Achievement	20.19	3.40	21.65	2.73
Safety	21.22	3.46	22.12	2.85
Comfort	37.44	4.27	39.31	3.45
Status	21.48	2.69	21.74	2.97
Altruism	21.48	3.55	22.18	3.07
Autonomy	19.26	2.08	20.95	2.14
EXPRESSED INTEREST	16.00	2.93	15.58	3.30

Table 7.18
Fort Lewis Pilot Test: AVOICE Scale Intercorrelations

VAR:	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
19I AVOICE MARKSMAN	88	44	15	02	27	40	18	12	24	28	70	40	72	73	49	28	19	51	48	40	56	26
20I AVOICE AGRICULTURE	44	88	32	43	40	35	20	35	34	38	50	54	54	53	43	39	28	43	52	41	60	42
21I AVOICE MATHEMATICS	15	32	88	56	58	62	58	62	59	54	32	18	17	26	53	70	71	15	34	42	16	56
22I AVOICE AESTHETICS	02	43	56	88	54	44	36	65	41	39	20	17	10	19	44	55	55	18	21	29	09	53
23I AVOICE LEADERSHIP	27	40	58	54	88	44	49	78	44	55	42	19	40	41	53	55	56	51	30	31	15	63
24I AVOICE ELECTRONIC COMMUN.	40	35	62	44	44	88	67	38	54	61	57	34	41	55	58	53	49	36	36	63	38	49
25I AVOICE AUTOMATED DATA PROC.	18	20	58	36	49	67	88	35	51	57	25	10	22	26	48	49	54	21	24	56	14	44
26I AVOICE TEACHING/COUNSELING	12	35	62	65	78	38	35	88	32	39	22	10	24	25	40	60	61	32	19	21	04	63
27I AVOICE DRAFTING	24	34	59	41	44	54	51	32	88	67	40	37	25	41	51	53	52	19	41	44	43	48
28I AVOICE AUDIOGRAPHICS	28	38	54	39	55	61	57	39	67	88	42	40	39	39	51	51	49	32	44	59	40	49
29I AVOICE ARMOR/CANNON	70	50	32	20	42	57	25	22	40	42	88	60	65	80	54	44	34	56	62	48	64	37
30I AVOICE VEHICLE OPERATOR	40	54	18	17	19	34	10	10	37	40	60	88	46	58	38	42	25	42	61	43	81	23
31I AVOICE OUTDOORS	72	54	17	10	40	41	22	24	25	39	65	46	88	72	40	19	10	61	58	41	61	27
32I AVOICE INFANTRY	73	53	26	19	41	55	26	25	41	39	80	58	72	88	58	39	29	70	45	34	62	39
33I AVOICE SCIENCE/CHEMICAL	49	43	53	44	53	58	48	40	51	51	54	38	40	58	88	51	49	49	42	54	39	56
34I AVOICE SUPPLY ADM.	28	39	70	55	55	53	49	60	53	51	44	42	19	39	51	88	90	27	36	47	37	67
35I AVOICE OFFICE ADM.	19	28	71	55	56	49	54	61	52	49	34	25	10	29	49	90	88	20	21	33	21	68
36I AVOICE LAW ENFORCEMENT	51	43	15	18	51	36	21	32	19	32	56	42	61	70	49	27	20	88	27	25	36	38
37I AVOICE MECHANICS	48	52	34	21	30	36	24	19	41	44	62	61	58	45	47	36	21	27	88	69	68	22
38I AVOICE ELECTRONICS	40	41	42	29	31	63	56	21	44	59	48	43	41	34	54	47	33	25	69	88	54	32
39I AVOICE HEAVY CONSTRUCTION	56	60	16	09	15	38	14	04	43	40	64	81	61	62	39	37	21	36	68	54	88	26
40I AVOICE MEDICAL	26	42	56	53	63	49	44	63	48	49	37	23	27	39	56	67	68	38	22	32	26	88
41I AVOICE FOOD SERVICE	13	37	52	59	36	36	36	45	35	19	21	23	06	23	31	60	60	12	20	28	20	48
42I AVOICE HOLLAND INVEST.	-08	-00	17	15	10	12	20	08	21	16	-04	-01	-08	04	18	19	24	04	-03	10	00	14
43I AVOICE HOLLAND CONVENT.	14	13	18	02	16	26	14	12	12	17	18	25	08	25	20	26	23	18	07	18	15	19
44I AVOICE HOLLAND ART.	15	20	31	33	29	22	23	30	32	30	18	19	06	22	25	42	43	13	13	19	10	29
45I AVOICE HOLLAND REAL.	33	23	-05	-15	04	10	-07	-10	17	12	29	32	31	36	16	04	-07	26	23	17	41	-08
46I AVOICE HOLLAND SOCIAL	12	16	21	19	23	11	17	34	15	13	06	00	00	10	21	28	32	11	03	12	-04	30
47I AVOICE HOLLAND ENTERPR.	15	05	22	08	27	16	15	20	29	27	15	11	08	20	28	21	17	19	10	14	04	24
48I AVOICE EXPRESSED INTEREST	29	11	11	12	19	18	18	10	11	09	29	10	24	34	21	11	09	36	19	14	11	12
49I AVOICE ABILITY UTIL.	-01	-01	03	13	23	-06	08	27	-10	-05	-08	-08	14	-01	03	01	06	07	-14	-10	-06	-05
50I AVOICE ACHIEVEMENT	-05	05	-02	28	16	-01	-06	29	-08	-07	-08	-07	18	-02	-09	-04	-00	08	-12	-19	-11	-05
----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
51I AVOICE ACTIVITY	15	33	02	13	22	12	11	20	-08	05	08	00	29	19	16	-06	05	27	-01	-06	08	02
52I AVOICE ADVANCEMENT	-00	-17	02	02	16	03	09	09	15	13	-08	-03	07	01	06	-06	-06	08	-06	-05	-10	-00
53I AVOICE AUTHORITY	15	12	05	05	29	-04	-01	23	08	-01	12	-13	10	06	08	07	08	04	04	-03	-01	07
54I AVOICE ORGANIZATION PSP	-04	-06	16	15	27	-05	03	21	-01	03	00	02	15	-06	10	13	07	07	15	00	-03	05
55I AVOICE COMPENSATION	-04	-06	19	16	19	07	15	30	02	01	-01	-03	06	04	09	08	11	10	-04	-08	-05	04
56I AVOICE CO-WORKERS	05	-03	10	03	22	12	12	17	-05	12	12	14	22	02	06	07	08	09	17	09	08	-01
57I AVOICE CREATIVITY	-07	08	02	08	07	01	-07	12	-06	05	-04	-00	11	-05	06	-10	-14	-05	05	-05	-04	03
58I AVOICE INDEPENDENCE	-06	-03	-33	-25	-31	-10	-14	-28	-11	-14	-00	12	-05	02	-05	-27	-24	05	-09	-07	13	-24
59I AVOICE MORAL VALUES	-17	05	21	22	09	02	09	25	-04	-03	-11	-09	-02	-21	-06	07	08	-13	13	07	-09	-12
60I AVOICE RECOGNITION	03	08	03	08	23	-02	02	21	11	12	03	11	06	-01	05	12	10	00	02	-01	12	12
61I AVOICE RESPONSIBILITY	-02	-05	-03	09	12	-01	-19	08	-03	09	01	-05	02	-05	-01	-00	-02	04	-07	-06	-10	04
62I AVOICE SECURITY	15	04	07	10	26	05	12	15	02	11	09	-01	34	17	09	-05	-04	33	01	-05	05	11
63I AVOICE SOCAIL SERVICE	22	44	49	57	63	36	34	63	40	35	37	23	35	32	41	50	52	37	39	24	22	55
64I AVOICE SOCIAL STATUS	02	19	27	41	50	22	22	54	19	29	06	-03	19	09	18	24	24	22	-02	05	-09	20
65I AVOICE SUPERVISION HR	-05	-10	10	13	17	07	08	19	-14	05	-18	-19	17	-11	01	-06	-05	01	-09	04	-15	-08
66I AVOICE SUPERVISION TECH	02	-06	20	10	33	17	29	30	15	01	-02	-17	09	11	19	12	17	05	-02	04	-14	19
67I AVOICE VARIETY	-18	-04	07	13	20	07	07	06	14	17	07	07	03	-04	-07	-05	-05	03	-01	-04	01	06
68I AVOICE WORKING COND.	-12	-13	-17	-10	02	-21	-04	08	-19	00	-21	-09	-03	-19	-16	-21	-19	-03	-13	-20	-12	-11

314

Table 7.18 (Continued)
Fort Lewis Pilot Test: AVOICE Scale Intercorrelations

	VAR.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
AVOICE MARKSMAN	19)	13	-08	14	15	33	12	15	29	-01	-05	15	-00	15	-04	-04	05	-07	-06	-17	03
AVOICE AGRICULTURE	20)	37	-00	13	20	23	16	05	11	-01	05	33	-17	12	-06	-06	-03	08	-02	05	08
AVOICE MATHEMATICS	21)	52	17	18	31	-05	21	22	11	03	-02	02	02	05	16	19	10	02	-33	21	03
AVOICE AESTHETICS	22)	59	15	02	33	-15	19	08	12	13	28	13	02	05	15	16	03	08	-25	22	08
AVOICE LEADERSHIP	23)	36	10	16	29	04	23	27	19	23	16	22	14	29	27	19	22	07	-31	09	23
AVOICE ELECTRONIC COMMUN.	24)	36	12	26	22	10	11	16	18	-06	-01	12	03	-04	-05	07	12	01	-10	02	-02
AVOICE AUTOMATED DATA PROC.	25)	36	20	14	23	-07	17	15	18	08	-06	11	09	-01	03	15	12	-07	-14	09	02
AVOICE TEACHING/COUNSELING	26)	45	08	12	30	-10	34	20	10	27	29	20	09	23	21	30	17	12	-28	25	21
AVOICE DRAFTING	27)	35	21	12	32	17	15	29	11	-10	-08	-08	15	08	-01	02	-05	-06	-11	-04	11
AVOICE AUDIOGRAPHICS	28)	19	16	12	30	12	13	27	09	-05	-07	05	13	-01	03	01	12	05	-14	-03	12
AVOICE ARMOR/CANNON	29)	21	-04	18	18	29	06	15	29	-08	-08	08	-08	12	00	-01	12	-04	-00	-11	03
AVOICE VEHICLE OPERATOR	30)	23	-01	25	19	32	00	11	10	-08	-07	00	-03	-13	02	-03	14	-00	12	-09	11
AVOICE OUTDOORS	31)	06	-08	08	06	31	00	08	24	14	18	29	07	10	15	06	22	11	-05	-02	06
AVOICE INFANTRY	32)	23	04	25	22	36	10	20	34	-01	-02	19	01	06	-06	04	02	-05	02	-21	-01
AVOICE SCIENCE/CHEMICAL	33)	31	18	20	25	16	21	28	21	03	-09	16	06	08	10	09	06	06	-05	-06	05
AVOICE SUPPLY ADM.	34)	60	19	26	42	04	28	21	11	01	-04	-06	-06	07	13	08	07	-10	-27	07	12
AVOICE OFFICE ADM.	35)	60	24	23	43	-07	32	17	09	06	-00	05	-06	08	07	11	08	-14	-24	08	10
AVOICE LAW ENFORCEMENT	36)	12	04	18	13	26	11	19	36	07	08	27	08	04	07	10	09	-05	05	-13	00
AVOICE MECHANICS	37)	20	-03	07	13	23	03	10	19	-14	-12	-01	-06	04	15	-04	17	05	-09	13	02
AVOICE ELECTRONICS	38)	25	10	15	19	17	12	14	14	-10	-19	-06	-05	-03	00	-05	09	-05	-07	07	-01
AVOICE HEAVY CONSTRUCTION	39)	20	00	15	10	41	-04	04	11	-06	-11	08	-10	-01	-03	-05	08	-04	13	-09	12
AVOICE MEDICAL	40)	48	14	19	29	-08	30	24	12	-05	-05	02	-00	07	05	04	-01	03	-24	-12	12
AVOICE FOOD SERVICE	41)	**	13	13	39	-05	10	11	10	03	06	04	-18	04	-02	04	-13	-14	-22	05	-05
AVOICE HOLLAND INVEST.	42)	13	**	17	22	21	22	25	19	03	02	14	-09	-03	-07	10	-06	06	06	-04	-14
AVOICE HOLLAND CONVENT.	43)	13	17	**	23	19	22	15	22	-06	-03	09	-11	-15	-22	00	-02	-17	17	-08	01
AVOICE HOLLAND ART.	44)	39	22	23	**	28	54	33	03	-08	-03	07	-15	-10	-07	03	04	-10	-11	06	-06
AVOICE HOLLAND REAL.	45)	-05	21	19	28	**	26	31	04	-01	-05	06	-10	02	-19	04	-09	-04	16	-16	-05
AVOICE HOLLAND SOCIAL	46)	10	22	22	54	26	**	44	-01	05	-08	12	-09	05	-14	08	-03	-02	08	04	02
AVOICE HOLLAND ENTERPR.	47)	11	25	15	33	31	44	**	07	-10	-14	-06	05	13	-08	09	-16	-00	-11	-16	-01
AVOICE EXPRESSED INTEREST	48)	10	19	22	03	04	-01	07	**	08	08	04	01	-07	-06	02	-05	-02	18	-14	-06
AVOICE ABILITY UTIL.	49)	03	03	-06	-08	-01	05	-10	08	**	36	42	18	09	28	32	28	09	03	13	24
AVOICE ACHIEVEMENT	50)	06	02	-03	-03	-05	-08	-14	08	36	**	31	09	14	14	31	33	12	-02	24	06
AVOICE ACTIVITY	51)	04	14	09	07	06	12	-06	04	42	31	**	-12	-03	10	09	26	-01	12	31	-01
AVOICE ADVANCEMENT	52)	-18	-09	-11	-15	-10	-09	05	01	18	09	-12	**	08	28	18	10	18	-13	-11	23
AVOICE AUTHORITY	53)	04	-03	-15	-10	02	05	13	-07	09	14	-03	08	**	09	11	00	21	-11	10	18
AVOICE ORGANIZATION P&P	54)	-02	-07	-22	-07	-19	-14	-08	-06	28	14	10	28	09	**	23	32	23	-25	26	17
AVOICE COMPENSATION	55)	04	10	00	03	04	08	09	02	32	31	09	18	11	23	**	13	11	01	29	24
AVOICE CO-WORKERS	56)	-13	-06	-02	04	-09	-03	-16	-05	28	33	26	10	00	32	13	**	13	-08	28	19
AVOICE CREATIVITY	57)	-14	06	-17	-10	-04	-02	-00	-02	09	12	-01	18	21	23	11	13	**	01	04	15
AVOICE INDEPENDENCE	58)	-22	06	17	-11	16	08	-11	18	03	-02	12	-13	-11	-25	01	-08	01	**	-20	-01
AVOICE MORAL VALUES	59)	08	-04	-08	06	-16	04	-16	-14	13	24	31	-11	10	26	29	28	04	-20	**	-01
AVOICE RECOGNITION	60)	-08	-14	01	-06	-05	02	-01	-06	24	06	-01	23	18	17	24	19	15	-01	-01	**
AVOICE RESPONSIBILITY	61)	-15	13	-08	00	00	10	13	17	17	16	14	03	09	15	-08	20	12	-04	-00	-01
AVOICE SECURITY	62)	-14	-17	-15	-17	-14	-08	-12	13	27	28	30	24	13	24	27	41	04	-01	18	27
AVOICE SOCAIL SERVICE	63)	47	05	10	26	-04	27	12	05	14	20	29	10	23	15	16	18	04	-15	22	14
AVOICE SOCIAL STATUS	64)	17	12	-03	11	00	14	06	-08	32	44	19	22	26	20	29	20	13	-21	25	07
AVOICE SUPERVISION HR	65)	-09	-08	-19	-03	-18	-09	-06	-08	35	31	27	31	06	38	24	43	16	-23	31	15
AVOICE SUPERVISION TECH	66)	09	03	05	06	-06	07	03	-03	22	12	10	29	15	19	27	07	11	-20	14	14
AVOICE VARIETY	67)	01	04	-11	-24	-09	-18	-09	01	11	08	-01	19	-03	30	-02	12	04	-03	-00	04
AVOICE WORKING COND.	68)	-29	-29	-08	-16	-27	01	-11	-11	23	26	04	37	05	17	21	27	23	18	12	34

Table 7.18 (Continued)

Fort Lewis Pilot Test: AVOICE Scale Intercorrelations

	VAR.	61	62	63	64	65	66	67	68
AVOICE MARKSMAN	191	-02	15	22	02	-05	02	-18	-12
AVOICE AGRICULTURE	201	-05	04	44	19	-10	-06	-04	-13
AVOICE MATHEMATICS	211	-03	07	49	27	10	20	07	-17
AVOICE AESTHETICS	221	09	10	57	41	13	10	13	-10
AVOICE LEADERSHIP	231	12	26	63	50	17	33	20	02
AVOICE ELECTRONIC COMMUN.	241	-01	05	36	22	07	17	07	-21
AVOICE AUTOMATED DATA PROC.	251	-19	12	34	22	08	29	07	-04
AVOICE TEACHING/COUNSELING	261	08	15	63	54	19	30	06	08
AVOICE DRAFTING	271	-03	02	40	19	-14	15	14	-19
AVOICE AUDIOGRAPHICS	281	09	11	35	29	05	01	17	00
AVOICE ARMOR/CANNON	291	01	09	37	06	-18	-02	07	-21
AVOICE VEHICLE OPERATOR	301	-05	-01	23	-03	-19	-17	07	-09
AVOICE OUTDOORS	311	02	36	35	19	17	09	03	-03
AVOICE INFANTRY	321	-05	17	32	09	-11	11	-04	-19
AVOICE SCIENCE/CHEMICAL	331	-01	09	41	18	01	19	-07	-16
AVOICE SUPPLY ADM.	341	-00	-05	50	24	-06	12	-05	-21
AVOICE OFFICE ADM.	351	-02	-04	52	24	-05	17	-05	-19
AVOICE LAW ENFORCEMENT	361	04	33	37	22	01	05	03	-03
AVOICE MECHANICS	371	-07	01	39	-02	-09	-02	-01	-13
AVOICE ELECTRONICS	381	-06	-05	24	05	04	04	-04	-20
AVOICE HEAVY CONSTRUCTION	391	-10	05	22	-09	-15	-14	01	-12
AVOICE MEDICAL	401	04	11	55	20	-08	19	06	-11
AVOICE FOOD SERVICE	411	-15	-14	42	17	-09	09	01	-29
AVOICE HOLLAND INVEST.	421	13	-17	05	12	-08	05	04	-29
AVOICE HOLLAND CONVENT.	431	-09	-15	10	-03	-19	05	-11	-08
AVOICE HOLLAND ART.	441	00	-17	26	11	-03	06	-24	-16
AVOICE HOLLAND REAL.	451	00	-14	-04	00	-18	-06	-09	-27
AVOICE HOLLAND SOCIAL	461	10	-08	27	14	-09	07	-18	01
AVOICE HOLLAND ENTERPR.	471	13	-12	12	06	-06	03	-09	-11
AVOICE EXPRESSED INTEREST	481	17	13	05	-08	-08	-03	01	-11
AVOICE ABILITY UTIL.	491	17	27	14	32	35	22	11	23
AVOICE ACHIEVEMENT	501	16	28	20	44	31	12	08	26
AVOICE ACTIVITY	511	14	30	29	19	27	10	-01	04
AVOICE ADVANCEMENT	521	03	24	10	22	31	29	19	37
AVOICE AUTHORITY	531	09	13	23	24	06	15	-03	05
AVOICE ORGANIZATION P&P	541	15	24	13	20	38	19	30	17
AVOICE COMPENSATION	551	-08	27	16	29	24	27	-02	21
AVOICE CO-WORKERS	561	20	41	18	20	43	07	12	27
AVOICE CREATIVITY	571	12	04	04	13	16	11	04	23
AVOICE INDEPENDENCE	581	-04	-01	-15	-21	-23	-20	-03	18
AVOICE MORAL VALUES	591	-00	18	22	25	31	14	-00	12
AVOICE RECOGNITION	601	-01	27	14	07	15	14	04	34
AVOICE RESPONSIBILITY	611	**	18	-02	09	18	-24	34	12
AVOICE SECURITY	621	18	**	29	17	36	14	08	31
AVOICE SOCAIL SERVICE	631	-02	29	**	35	13	24	-02	-05
AVOICE SOCIAL STATUS	641	09	17	35	**	20	23	22	13
AVOICE SUPERVISION HR	651	18	36	13	20	**	23	05	26
AVOICE SUPERVISION TECH	661	-24	14	24	23	23	**	-05	01
AVOICE VARIETY	671	34	08	-02	22	05	-03	**	12
AVOICE WORKING COND.	681	12	31	-05	13	26	01	12	**

SUMMARY

The two non-cognitive inventories of the Pilot Trial Battery, the ABLE and the AVOICE, are designed to measure a total of 20 constructs plus response validity scale and expressed interests categories. The ABLE assesses six temperament constructs and the Physical Condition construct through 11 scales, and also includes four response validity scales. The AVOICE measures six Holland interests constructs, six Organizational Environment constructs, and Expressed Interests through 31 scales. Altogether, the 46 scales of the two inventories included approximately 600 items during the pilot testing phase--291 ABLE items and 306 AVOICE items for the Fort Campbell version, and 268 ABLE items and 306 AVOICE items for the Fort Lewis version.

Evaluation and revision of the inventories took place in three steps. First, each was subjected to editorial review by both PDRI and ARI prior to any pilot testing. This review resulted in nearly 200 wording changes and the deletion of 17 items. The majority of these changes applied to ABLE.

The second stage of evaluation took place after the Fort Campbell pilot testing. Feedback from the soldiers taking the inventory and data analysis of the results (e.g., item-total correlations, item response distributions) were used to refine the inventories. Twenty-three ABLE items were deleted and 173 ABLE items were revised; no AVOICE items were deleted and 20 AVOICE items were revised.

In the third stage of evaluation, after the Fort Lewis pilot testing, far fewer changes were made. One ABLE item was deleted, 20 ABLE items were revised, and no changes were made to the AVOICE. Throughout the evaluation process, it is likely that the AVOICE was less subject to revision because it uses a common response format for all items, whereas the response options for ABLE items differ by item.

The psychometric data obtained with both inventories seemed highly satisfactory; the scales were shown to be reliable and appeared to be measuring the constructs intended. Sample sizes in these administrations were fairly small (Fort Campbell N = 52 and 55, ABLE and AVOICE, respectively; Fort Lewis N = 106 and 114, ABLE and AVOICE, respectively), but results were similar in both samples.

Chapter 7 References

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible applications. *The Journal of Psychology, 112*, 169-193.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook, Volume II: Research applications*. Minneapolis: University of Minnesota Press.

Dolliver, R. H. (1969). Strong Vocational Interest Blank versus expressed vocational interests: A review. *Psychological Bulletin, 72*, 95-107.

Gough, H. G. (1975). *Manual for the California Psychological Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1966). *The psychology of vocational choice.* Waltham, MA: Blaisdell.

Hough, L. M., et al. *Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance.* ARI Research Note in preparation.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

CHAPTER 8

NON-COGNITIVE MEASURES: FIELD TESTS

Leaetta M. Hough, Matthew K. McGue, Janis S. Houston, and Elaine D. Pulakos

In this chapter we describe the field tests of the non-cognitive measures in the Pilot Trial Battery, the ABLE and the AVOICE, whose development was described in Chapter 7. Portions of this chapter are drawn from Hough, Barge, Houston, McGue, and Kamp (1985).

We first discuss the results of the Fort Knox field test in September 1984, the general procedures for which were described in Chapter 2. We also discuss here the procedures and results of the field testing done at Fort Bragg, where the ABLE and AVOICE were administered to soldiers under several experimental conditions, in order to estimate the extent to which scores on these inventories could be "faked" when individuals are instructed to do so. We also describe, in the context of this "fakability" study, the procedures and results of the ABLE and AVOICE administration to recruits at the Military Entrance Processing Station (MEPS) at Minneapolis.

Figures 8.1 and 8.2 list the entire set of scales, by construct, contained in the Fort Knox version of the ABLE and AVOICE, respectively. Chapter 7 presented a complete description of each of these constructs and scales, with sample items, and the two inventories themselves in the form administered at Fort Knox, may be found in Appendix G.

Construct	Scale
Adjustment	Emotional Stability
Dependability	Nondelinquency Traditional Values Conscientiousness
Achievement	Work Orientation Self-Esteem
Physical Condition	Physical Condition
Leadership (Potency)	Dominance Energy Level
Locus of Control	Internal Control
Agreeableness/Likeability	Cooperativeness
Response Validity Scales	Non-Random Response Unlikely Virtues (Social Desirability) Poor Impression Self-Knowledge

Figure 8.1 ABLE scales organized by construct.

Realistic Interests

- Basic Interest Item
- Mechanics
- Heavy Construction
- Electronics
- Electronic Communication
- Drafting
- Law Enforcement
- Audiographics
- Agriculture
- Outdoors
- Marksman
- Infantry
- Armor/Cannon
- Vehicle Operator
- Adventure

Conventional Interests

- Basic Interest Item
- Office Administration
- Supply Administration
- Food Service

Social Interests

- Basic Interest Item
- Teaching/Counseling

Investigative Interests

- Basic Interest Item
- Medical Services
- Mathematics
- Science/Chemical
- Automated Data Processing

Enterprising Interests

- Basic Interest Item
- Leadership

Artistic Interests

- Basic Interest Item
- Aesthetics

Organizational Climate/Environment Preferences

- Achievement Preferences
- Safety Preferences
- Comfort Preferences
- Status Preferences
- Altruism Preferences
- Autonomy Preferences

Expressed Interests

- Expressed Interests

Figure 8.2 AVOICE scales organized by construct.

ANALYSIS OF DATA FROM FIELD TEST ADMINISTRATION

Results of Data Quality Screening

In Table 8.1, the data screening results are presented for the Fort Knox field test. A total of 290 soldiers completed the ABLE and 287 soldiers completed the AVOICE. After deletion of inventories with greater than 10 percent missing data for both inventories, and deletion of those ABLEs where scores on the Non-Random Response Scale (NRRS) were less than six, a total of 276 ABLEs and 270 AVOICEs were available for analysis.

Recall from Chapter 2 that portions of the Pilot Trial Battery were re-administered to soldiers two weeks after the first administration. As can be seen in Table 8.1, the total number of "Time 2" ABLE and AVOICE inventories, after the data quality screens had been applied, was 109 and 127, respectively.

Mean Scores and Reliability Estimates

Summary statistics for the non-cognitive measures are presented in Tables 8.2, 8.3, and 8.4. Several things are noteworthy in Table 8.2. All the ABLE content scales show adequate score variances (SD ranges from 5.25 to 8.27) and the alpha coefficients are acceptable to excellent in value (median = .84, range = .70 to .87). In passing, we point out that there was no particular technical reason for computing alpha coefficients on the field test data rather than Hoyt coefficients as was done for the pilot data test (see Chapter 7). Both procedures provide conceptually identical estimates of internal consistency reliability and provide nearly identical mathematical results. Other work on Project A was using the alpha coefficient procedure, so we decided to use the same procedure for the sake of greater project-wide consistency. The test-retest coefficients are all at or greater than acceptable levels (median = .79, range = .68 to .83), and in most cases are near the same value as the alphas, indicating excellent stability for these scale scores.

The response validity scales have score variances as expected. Unlikely Virtues and Self-Knowledge scores are nearly normally distributed with somewhat less variance than the content scales, but still on an acceptable level. The Non-Random Response and Poor Impression scales show markedly skewed distributions as would be expected for subjects responding attentively and honestly. The alphas for these scales are a bit lower than for the content scales, again as expected. The test-retest coefficients are also a bit lower, especially for Non-Random Response. However, the variance is small on this scale (again, as it should be) and the distribution is skewed, so even small changes in responses can have a large effect on this coefficient.

Table 8.3 shows more detail about the test-retest results for the ABLE. The results for the content scales, which are the most important scales in terms of predicting job performance and other criteria, are remarkable for their consistency. There was virtually no change in mean scores between the two administrations, and the effect sizes are very small.

The response validity scales appear to be more sensitive to changes

Table 8.1

Fort Knox Field Test: Data Quality Screen Results

	Fort Knox Time 1		Fort Knox Time 2	
Total N at Sessions	303		258	
ABLE				
N taking this inventory	290		128	
Number deleted with Overall Missing Data Screen (>10%, or 27 items)	9	(3%)	7	(5%)
Number deleted with NRRS[a] Screen (<6 "correct" out of 8)	5	(2%)	12	(9%)
N usable ABLEs	276	(95%)	109	(85%)
AVOICE				
N taking this inventory	287		130	
Number deleted with Overall Missing Data Screen (>10%, or 31 items)	17	(6%)	3	(2%)
N usable AVOICEs	270	(94%)	127	(98%)

---

[a] Non-Random Response Scale.

Table 8.2

Fort Knox Field Test: ABLE Scale Score Characteristics
(N = 276 except where otherwise noted)

Scale	Number of Items	Mean Time 1	SD	Alpha	Test-Retest[a] r	Median Item-Scale r
Content Scales						
Emotional Stability	29	64.9	8.27	.86	.68	.44
Self-Esteem	15	35.1	5.25	.83	.81	.54
Cooperativeness	24	54.1	6.09	.77	.69	.42
Conscientiousness	21	48.9	5.90	.81	.73	.43
Nondelinquency	24	55.4	7.23	.84	.81	.46
Traditional Values	16	37.2	4.60	.70	.74	.45
Work Orientation	27	61.2	7.93	.85	.80	.47
Internal Control	21	50.3	6.14	.79	.75	.43
Energy Level	25	57.1	7.11	.85	.79	.47
Dominance	16	35.5	6.13	.86	.83	.56
Physical Condition	9	31.1	7.53	.87	.81	.72
Response Validity Scales						
Unlikely Virtues	12	16.6	3.39	.68	.62	.53
Self-Knowledge	13	29.6	3.54	.62	.71	.41
Non-Random Response[b]	8	7.7	.71	.56	.37	.45
Poor Impression	24	1.5	1.86	.61	.56	.33

[a] N=109 for Test-Retest correlations. Test-Retest interval was two weeks.

[b] N=281. Statistics reported for this scale are based on sample edited for overall Missing Data only. "Passing" score on Non-Random Response Scale ≤ 6.

Table 8.3

Fort Knox Field Test: ABLE Test-Retest Results[a]

Scale	Mean Time 1 (N = 276)	Mean Time 2 (N = 109)	Effect Size[c]
Content Scales			
Emotional Stability	64.9	65.1	.02
Self-Esteem	35.1	34.8	-.05
Cooperativeness	54.1	54.3	.04
Conscientiousness	48.9	48.3	-.10
Nondelinquency	55.4	55.6	.02
Traditional Values	37.2	37.9	.15
Work Orientation	61.2	60.7	-.07
Internal Control	50.3	50.2	-.01
Energy Level	57.1	57.0	-.01
Dominance	35.5	34.9	-.09
Physical Condition	31.1	30.4	-.09
Response Validity Scales			
Unlikely Virtues	16.6	17.5	.27
Self-Knowledge	29.6	29.0	-.18
Non-Random Response[b]	7.7	7.2	-.65
Poor Impression	1.5	1.2	-.18

---

[a] Test-Retest interval was two weeks.

[b] Based on sample edited for missing data only; $N_1 = 281$ and $N_2 = 121$.

[c] Effect Size = (Mean Time 1 - Mean Time 2)/SD Time 1

Table 8.4

Fort Knox Field Test: AVOICE Scale Score Characteristics
(N = 270 except where otherwise noted)

Scale	Number of Items	Mean	SD	Alpha	Test-Retest[a] r	Median Item-Scale r
Marksman	5	15.8	4.37	.79	.77	.75
Agriculture	5	14.1	3.99	.68	.69	.70
Mathematics	5	15.1	4.37	.82	.76	.79
Aesthetics	5	14.3	4.17	.77	.72	.74
Leadership	6	20.3	4.70	.81	.56	.74
Electronic Communication	7	21.1	5.73	.92	.78	.72
Automated Data Processing	7	23.4	6.56	.88	.81	.81
Teaching/Counseling	7	22.8	5.53	.82	.73	.73
Drafting	7	21.5	6.12	.85	.74	.77
Audiographics	7	23.8	5.68	.82	.76	.70
Armor/Cannon	8	22.4	6.57	.83	.74	.69
Vehicle/Equipment Operator	10	28.1	7.79	.86	.69	.70
Outdoors	9	31.7	6.41	.79	.69	.66
Infantry	10	29.1	7.13	.81	.78	.65
Science/Chemical Operations	11	29.4	8.93	.89	.79	.71
Supply Administration	13	35.0	10.44	.92	.82	.75
Office Administration	16	45.2	13.20	.94	.86	.73
Law Enforcement	16	48.1	11.84	.88	.78	.63
Mechanics	16	50.0	14.68	.95	.80	.80
Electronics	20	60.0	17.48	.96	.74	.77
Heavy Construction/Combat	23	65.8	17.90	.94	.76	.70
Medical Services	24	68.5	18.79	.95	.84	.69
Food Service	17	48.2	11.16	.89	.71	.64

---

[a] N=127 for Test-Retest correlations.

due to time or due to a second administration. The change in mean scores is greater than for the content scales and the effect sizes are somewhat larger. Still, the changes are not large except for the Non-Random Response score. The change in this mean score indicates that more subjects responded less attentively the second time around, which is perhaps not surprising. We point out that the Non-Random Response Scale did "catch" this phenomenon, exactly as it was supposed to, and roughly four times as many subjects "failed" this scale on the second administration as did on the first (2 percent vs. 9 percent, see Table 8.1). Overall we find these results reassuring with respect to the way the content and response validity scales were designed to function.

Table 8.4 shows that the AVOICE scales are also functioning well. Scale score statistics show adequate variance (SD ranges from 3.99, for a scale with a possible score range of 5-25, to 18.79, for a scale with a possible score range from 24-12). Alpha coefficients vary from .68 to .96 with a median of .86, with the lower values occurring for the scales with fewer items, as would be expected. The median item-total scale score correlations are all very high (.60s to .70s), also indicating good internal consistency. Finally, the test-retest coefficients are also acceptable to excellent in value (median value = .76, range from .56 to .86).

The results shown in Tables 8.2, 8.3, and 8.4 and discussed above lead to the conclusion that the non-cognitive scales are very sound with regard to basic psychometric criteria of sufficient score variance and distribution, internal consistency, and stability.

Uniqueness Estimates for Non-Cognitive Measures

Scales on both the ABLE and the AVOICE were examined for their potential for providing incremental validity to the predictor battery. Uniqueness estimates were computed identically to those described for the cognitive measures in Chapter 4, by subtracting the squared multiple regression of a set of tests (e.g., the ASVAB) from the reliability estimate for the test of interest ($U^2 = R_{xx}-R^2$). Uniqueness is, then, the amount of reliable variance for a test not shared with the tests against which it has been regressed.

Tables 8.5 and 8.6 present the uniqueness estimates for the ABLE and AVOICE scales, respectively, when regressed against the ASVAB. The median U^2 for the ABLE is .80, and ranges from .69 to .87, indicating that the ABLE overlaps very little with the ASVAB. The median estimate of uniqueness for the AVOICE is .81 and ranges from .59 to .95, indicating that the AVOICE also overlaps very little with the ASVAB.

Table 8.7 contains a summary of the correlations between the ABLE and the AVOICE, and the other measures in the Pilot Trial Battery. As can be seen here, the ABLE and AVOICE share very little variance with the cognitive and psychomotor tests in the Pilot Trial Battery.

Factor Analysis of ABLE and AVOICE Scales

The ABLE content scales and the AVOICE scales were separately factor analyzed, and, in both cases, a two-factor solution appeared to best summarize the data. Table 8.8 shows the factor loading matrix for the ABLE

Table 8.5

Uniqueness Estimates for 11 ABLE Scales in the Pilot Trial Battery Against Other ABLE Scores and Against ASVAB

Scale	Number of Items	Alpha (N=276)	Test-Retest r (N=109)	ABLE Adj R^2 (N=207)	ASVAB Adj R^2 (N=183)	ASVAB U^2 Using Alpha (N=183)	ASVAB U^2 Using T-R (N=183)
Emotional Stability	29	.86	.68	.52	.05	.81	.63
Self-Esteem	15	.83	.81	.70	.03	.80	.78
Cooperativeness	24	.77	.69	.54	.00	.77	.69
Conscientiousness	21	.81	.73	.64	.03	.78	.70
Nondelinquency	24	.84	.81	.63	.02	.82	.79
Traditional Values	16	.70	.74	.50	.01	.69	.73
Work Orientation	27	.85	.80	.71	.03	.82	.77
Internal Control	21	.79	.75	.48	.04	.75	.71
Energy Level	25	.85	.79	.72	.05	.80	.74
Dominance	16	.86	.83	.50	.00	.86	.83
Physical Condition	9	.87	.81	.11	.00	.87	.81

Table 8.6

Uniqueness Estimates for 24 AVOICE Scales in the Pilot Trial Battery Against ASVAB

Scale	Number of Items	Alpha (N=270)	Test-Retest r (N=127)	ASVAB Adj R^2 (N=149)	ASVAB U^2 Using Alpha (N=149)	ASVAB U^2 Using T-R (N=149)
Marksman	5	.79	.77	.20	.59	.57
Agriculture	5	.68	.69	.06	.62	.63
Mathematics	5	.82	.76	.02	.80	.74
Aesthetics	5	.77	.72	.08	.69	.64
Leadership	6	.81	.56	.00	.81	.56
Electronic Communication	7	.92	.78	.01	.91	.77
Automated Data Processing	7	.88	.81	.00	.88	.81
Teaching/Counseling	7	.82	.73	.00	.82	.73
Drafting	7	.85	.74	.07	.78	.67
Audiographics	7	.82	.76	.00	.82	.76
Armor/Cannon	8	.83	.74	.11	.72	.63
Vehicle/Equipment Operator	10	.86	.69	.14	.72	.55
Outdoors	9	.79	.69	.16	.63	.53
Infantry	10	.81	.78	.13	.68	.65
Science/Chemical Operations	11	.89	.79	.01	.88	.78
Supply Administration	13	.92	.82	.00	.92	.82
Office Administration	16	.94	.86	.03	.91	.83
Law Enforcement	16	.88	.78	.02	.86	.76
Mechanics	16	.95	.80	.32	.63	.48
Electronics	20	.96	.74	.14	.82	.60
Heavy Construction/Combat	23	.94	.76	.21	.73	.55
Medical Services	24	.95	.84	.00	.95	.84
Food Service	17	.89	.71	.02	.87	.69
Adventure	14	.96	.86	.26	.70	.60

Table 8.7

Summary of Overlap of Non-Cognitive Measures With Other Pilot Trial Battery Measures

---

1. Between ABLE and PTB Cognitive Paper-and-Pencil Tests:
   - Only 19%, 29 of 150 correlations, are significant at $p \leq .05$.
   - The highest correlation is .23.

2. Between ABLE and PTB Computer-Administered Measures:
   - Only 17%, 48 of 285 correlations, are significant at $p \leq .05$.
   - The highest correlation is .24.

3. Between AVOICE and PTB Cognitive Paper-and-Pencil Tests:
   - Only 36%, 128 of 130 correlations, are significant at $p \leq .05$.
   - The highest correlation is .32.

4. Between AVOICE and PTB Computer-Administered Measures:
   - Only 15%, 105 Of 684 corelations, are significant at $p \leq .05$.
   - The highest correlation is .30.

Table 8.8

Fort Knox Field Test: ABLE Factor Analysis[a]
(N = 276)

	I Personal Impact	II Dependability	h^2
Self-Esteem	.80	.30	.73
Energy Level	.73	.46	.74
Dominance (Leadership)	.72	.13	.54
Emotional Stability	.67	.26	.52
Work Orientation	.67	.51	.71
Nondelinquency	.20	.81	.70
Traditional Values	.19	.73	.57
Conscientiousness	.39	.72	.67
Cooperativeness	.46	.60	.57
Internal Control	.44	.50	.44
			6.19

Note: h^2 = communality, the sum of squared factor loadings for a variable.

[a] Principal factor analysis, varimax rotation.

content scales. Note first that the communalities for the scales are fairly high, indicating that the scales do share substantial common variance.

The first factor was labeled Personal Impact since the scales loading on the factor, in concert, suggest that persons scoring high on the factor would have high self-esteem, exhibit a high level of energy, could exert leadership, would appear emotionally stable, and would be work oriented. Note that two of the scales loading highest on this factor do have substantial loadings on the second factor--Energy Level (.46) and Work Orientation (.51). Also, three of the scales loading highest on the second factor had substantial loadings here--Cooperativeness (.46), Internal Control (.44), and Conscientiousness (.39).

The second factor was named Dependability. Scale loadings for this factor suggest that a high scorer on this factor would be a strong rule abider, a believer in traditional societal values, show conscientiousness, be cooperative, and believe that life's circumstances were largely under an individual's control. Again, keep in mind the scales that show high loadings on both factors (as noted in the above paragraph).

This two-factor solution seems to us to make good intuitive sense for characterizing soldiers as well as possessing a fair amount of practical appeal. Being able to identify soldiers with high personal impact or leadership potential and a high degree of dependability would seem to be a potentially valuable contribution.

The solution found in these field test data differs from the pilot test solution primarily in the number of factors that characterize the best solution. Two factors were viewed as best here, whereas a larger number of factors were viewed as best in those solutions (see Table 7.11). The most probable reason for this difference is the difference in the two samples. The field test results are based on a sample roughly two and one-half times as large and is probably a more representative sample in terms of diversity of MOS as well. Therefore, we think the field test data are "better" data to interpret.

Table 8.9 shows the results for the factor analysis of the AVOICE. The scale communalities for this AVOICE solution are a bit lower than those for the ABLE, but still do indicate a substantial amount of common variance for the set of scales. (Sixty-two percent of the total ABLE scale variance is in common compared to 54 percent for the AVOICE).

The two factors found here were named Combat Support and Combat-Related. The former is defined largely by scales that have to do with jobs or services that support the actual combat specialties, while the latter is defined by scales that, for the most part, are much more related to specialties that engage directly in combat.

Also, as found with the ABLE, several scales show substantial loadings on both factors. Most of these occur for scales loading highest on the first factor, and include Science/Chemical Operations (.43 on second factor), Electronic Communication (.36), Leadership (.35), and Drafting (.34). Only one scale loading highest on the second factor has a substantial loading on the first factor, Electronics (.45).

Table 8.9

Fort Knox Field Test: AVOICE Factor Analysis[a]
(N = 270)

Scale	I Combat Support[b]	II Combat-Related[c]	h^2
Office Administration	.85	-.13	.73
Supply Administration	.78	.11	.62
Teaching/Counseling	.76	.11	.59
Mathematics	.74	.09	.55
Medical Services	.73	.18	.57
Automated Data Processing	.71	.10	.51
Audiographics	.64	.17	.44
Electronic Communication	.64	.36	.54
Science/Chemical Operations	.61	.43	.55
Aesthetics	.61	.04	.37
Leadership	.58	.35	.46
Food Service	.54	.19	.33
Drafting	.54	.34	.41
Infantry	.10	.85	.74
Armor/Cannon	.13	.84	.73
Heavy Construction/Combat	.17	.84	.73
Outdoors	.02	.74	.55
Mechanics	.17	.74	.58
Marksman	.05	.73	.54
Vehicle/Equipment Operator	.17	.73	.56
Agriculture	.18	.64	.44
Law Enforcement	.27	.61	.44
Electronics	.45	.57	.52
			12.49

Note: h^2 = communality, the sum of squared factor loadings for a variable.

[a] Principal factor analysis, varimax rotation.

[b] Conventional, Social, Investigative, Enterprising, Artistic constructs.

[c] Realistic construct.

The remarks made above about the comparison of ABLE factor analyses of the pilot and field test data apply equally here. Again, we think the field test data are probably the better set of results in terms of the representativeness of the samples.

Finally, as with the ABLE, we think the two-factor AVOICE solution makes good intuitive sense and has practical appeal. It would seem to be helpful to be able to characterize applicants as having interests primarily in the combat MOS or in MOS supporting combat specialties, perhaps even at the point of recruitment as opposed to the selection or in-processing point.

FAKABILITY INVESTIGATIONS

As discussed previously, in addition to the content scales, there were four response validity scales on the ABLE: Non-Random Response, Unlikely Virtues (Social Desirability), Poor Impression, and Self-Knowledge. An investigation was undertaken, including an experiment, on intentional distortion (faking) of responses. Data were gathered for this study from (1) soldiers instructed, at different times, to distort their responses and to be honest (experimental data gathered at Fort Bragg); (2) soldiers who were simply responding to the ABLE and AVOICE with no particular directions (data gathered at Fort Knox, in another type of "honest" condition); and (3) recently sworn-in Army recruits at the Minneapolis Military Entrance Processing Station (MEPS).

Purposes of the Faking Study

The purposes of the faking study were to determine:

- The extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so. (Compare data from Fort Bragg faking conditions with Fort Bragg and Fort Knox honest conditions.)
- The extent to which the ABLE response validity scales detect such intentional distortion. (Compare response validity scales in Fort Bragg honest and faking conditions.)
- The extent to which ABLE validity scales can be used to correct or adjust scores for intentional distortion.
- The extent to which distortion might be a problem in an applicant settings. (Compare MEPS data with Fort Bragg and Fort Knox data.)

The participants in the experimental group were 425 enlisted soldiers in the 82nd Airborne brigade at Fort Bragg in September 1984. Comparison samples were new recruits at a MEPS, in an approximation of an applicant setting, (N = 126) and Fort Knox soldiers described earlier (N = 276).

Procedure and Design

Four faking conditions were created:

- Fake Good on the ABLE
- Fake Bad on the ABLE
- Fake Combat on the AVOICE
- Fake Noncombat on the AVOICE

Two honest conditions were created:

- Honest on the ABLE
- Honest on the AVOICE

The significant parts of the instructions for the six conditions were as follows:

- ABLE - Fake Good

  Imagine you are at the Military Entrance Processing Station (MEPS) and you want to join the Army. Describe yourself in a way that you think will ensure that the Army selects you.

- ABLE - Fake Bad

  Imagine you are at the Military Entrance Processing Station (MEPS) and you do not want to join the Army. Describe yourself in a way that you think will ensure that the Army does not select you.

- ABLE - Honest

  You are to describe yourself as you really are.

- AVOICE - Fake Combat

  Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way that you think will ensure that you are placed in an occupation in which you are likely to be exposed to combat during a wartime situation.

- AVOICE - Fake Noncombat

  Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way you think will ensure that you are placed in an occupation in which you are unlikely to be exposed to combat during a wartime situation.

- AVOICE - Honest

  You are to describe yourself as you really are.

The design was repeated measures with faking and honest conditions counter-balanced. Thus, approximately half the experimental group, 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121) completed the inventories honestly in the afternoon and faked in the morning.

The experimental design and the numbers of soldiers from whom we gathered the intentional faking data appear in Table 8.10. In summary, a 2 x 2 x 2 fixed-factor, completely crossed experimental design was used. The within-subjects factor, called "Fake," consisted of two levels (honest responses and faked responses). The first between-subjects factor, called "Set," consisted of the following two levels: Fake Good (for the ABLE)/ Want Combat (for the AVOICE) and Fake Bad (for the ABLE)/Do Not Want Combat (for the AVOICE). Order was manipulated in the second between-subjects factor such that the following two levels were produced: faked responses before honest responses, and honest responses before faked responses.

Table 8.10

Faking Experiment, ABLE and AVOICE: Fort Bragg

AVOICE/ABLE COUNTS

Monday				
AM:	Honest AVOICE	N=64	62	Complete Sets
	Honest ABLE			
PM:	Fake Combat AVOICE	N=62		
	Fake Good ABLE			
Tuesday				
AM:	Honest AVOICE	N=62	62	Complete Sets
	Honest ABLE			
PM:	Fake Noncombat AVOICE	N=62		
	Fake Bad ABLE			
Wednesday				
AM:	Fake Combat AVOICE	N=63	61	Complete Sets
	Fake Good ABLE			
PM:	Honest AVOICE	N=61		
	Honest ABLE			
Thursday				
AM:	Fake Noncombat AVOICE	N=61	60	Complete Sets
	Fake Bad ABLE			
PM:	Honest AVOICE	N=60		
	Honest ABLE			

Faking Study Results - Temperament Inventory

We performed a multivariate analysis of variance (MANOVA) on the experimental data from Fort Bragg. Table 8.11 shows the findings for the interactions, the sources of variance most relevant to the question of whether soldiers can or cannot intentionally distort their responses.

As can be seen, all the Fake x Set interactions are significant, indicating that soldiers can, when instructed to do so, distort their responses.

Table 8.11 also shows that, for the Fake x Set x Order interaction effect, the overall test of significance is statistically significant for the response validity scales and marginally significant for the content scales. These results indicate that the order of experimental conditions in which the participant completed the ABLE affected the results. Table 8.12 shows in greater detail the effects of intentional distortion; it shows the mean scores for the various experimental conditions for the content scales. This table and the remaining tables showing Fort Bragg ABLE results report the values for the soldier responses on the first administration of the particular condition. For example, the mean value of 66.1 for Emotional Stability in the Honest First column of Table 8.12 was computed on 120 soldiers who completed the ABLE under the Honest condition before they completed the ABLE under a Fake condition (either Good or Bad). Similarly, the mean value of 70.3 for Emotional Stability in the Fake Good First column of Table 8.12 was computed on 54 soldiers who completed the ABLE under the Fake Good condition before they completed the ABLE under the Honest condition.

In general, Table 8.12 shows scores are higher on all the content scales when subjects are instructed to fake good (about .5 SD on average), and, to a much greater extent, scores are lower on the content scales when subjects are instructed to fake bad (about 2 SDs on average).

Another research question was the extent to which our response validity scales detected intentional distortion. As can be seen in Table 8.13, the response validity scale Unlikely Virtues (Social Desirability) detects Faking Good on the ABLE; the response validity scales Non-Random Response, Poor Impression, and Self-Knowledge detect Faking Bad. According to these data, the soldiers responded more randomly, created a poorer impression, and reported that they knew themselves less well when told to describe themselves in a way that would increase the likelihood that they would not be accepted into the Army.

We also examined the extent to which we could use the response validity scales Unlikely Virtues (Social Desirability) and Poor Impression to adjust ABLE content scale scores for Faking Good and Faking Bad. We regressed out Social Desirability from the content scales in the Fake Good condition and Poor Impression from the content scales in the Fake Bad condition. Table 8.14 shows the adjusted mean differences in content scales after regressing out Social Desirability and Poor Impression. Comparing these differences to the unadjusted differences shown in Table 8.12 clearly shows that these response validity scales can be used to adjust content scales. However, two important unknowns remain: Do the adjustment

Table 8.11

Fakability Study, MANOVA Results for ABLE Scales: Fort Bragg

	Interactions	
Type and Name of Scale	Fake x Set	Fake x Set x Order
Response Validity Scales[a]		
Overall	S	S
Unlikely Virtues (Social Desirability)	S	S
Self-Knowledge	S	NS
Non-Random Response	S	NS
Poor Impression	S	NS
Content Scales[b]		
Overall	S	NS*
Emotional Stability	S	---
Self-Esteem	S	---
Cooperativeness	S	---
Conscientiousness	S	---
Nondelinquency	S	---
Traditional Values	S	---
Work Orientation	S	---
Internal Control	S	---
Energy Level	S	---
Dominance (Leadership)	S	---

Note: S = significant, $p \leq .01$.
NS = nonsignificant, $p > .01$.
* = marginally significant, $.05 \leq p > .01$.

[a] Sample size for Response Validity Scales is 219.

[b] Sample size for Content Scales is 208.

Table 8.12

Honesty and Faking Effects, ABLE Content Scales: Fort Bragg

	Honest First[a]			Fake Good First[a]			Fake Bad First[a]			Estimated Effect Size	Estimated Effect Size
Scale	**N**	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**	**Honest vs. Good**	**Honest vs. Bad**
Emotional Stability	120	66.1	7.8	54	70.3	10.2	54	50.1	10.8	-.49	1.81
Self-Esteem	115	34.8	4.7	54	38.2	5.4	54	22.2	5.8	-.69	2.48
Cooperativeness	121	53.2	6.3	54	55.5	8.8	54	36.7	10.4	-.32	2.12
Conscientiousness	116	46.3	5.8	54	49.6	8.4	54	31.7	8.7	-.49	2.13
Nondelinquency	115	53.1	6.2	54	54.8	10.2	54	36.8	9.6	-.22	2.19
Traditional Values	116	36.7	4.6	54	38.7	6.5	54	23.6	6.1	-.38	2.56
Work Orientation	120	59.3	7.6	54	64.7	10.3	54	40.8	11.7	-.63	2.04
Internal Control	115	49.5	6.3	54	50.9	8.2	54	35.6	8.9	-.20	1.92
Energy Level	116	57.5	6.9	54	61.4	9.1	54	37.9	9.9	-.51	2.46
Dominance (Leadership)	116	35.6	5.6	54	40.3	5.6	54	24.5	6.6	-.84	1.87
Physical Condition	116	33.0	7.4	54	35.4	7.7	54	18.3	8.6	-.32	1.88

---

[a] Mean scores are based on persons who responded to this condition first.

Table 8.13

Honesty and Faking Effects, ABLE Response Validity Scales: Fort Bragg

	Honest First[a]			Fake Good First[a]			Fake Bad First[a]			Effect Size Honest vs. Fake Good	Effect Size Honest vs. Fake Bad
ABLE Response Validity Scale	N	M	SD	N	M	SD	N	M	SD		
Unlikely Virtues (Social Desirability)	109	15.8	3.1	57	20.1	5.8	56	17.8	4.8	$\boxed{-1.02}$	-.53
Self-Knowledge	109	29.6	3.6	57	29.7	4.1	56	21.8	5.2	-.03	$\boxed{1.85}$
Non-Random Response	109	7.6	1.0	57	7.0	1.8	56	2.8	2.2	.45	$\boxed{3.16}$
Poor Impression	109	1.5	2.1	57	1.7	2.2	56	14.6	7.9	-.09	$\boxed{-2.67}$

[a] Values are based on the sample that completed the questionnaires under the condition of interest first.

Table 8.14

Effects of Regressing Out Two Response Validity Scales (Social Desirability and Poor Impression) on Faking Condition, ABLE Content Scale Scores: Fort Bragg

	Fake Good		Fake Bad	
Content Scales	Adjusted Standardized Mean Difference[a]	Correlation with Social Desirability[b]	Adjusted Standardized Mean Difference[a]	Correlation with Poor Impression[b]
Emotional Stability	-.14	.14	-.14	-.41
Self-Esteem	-.64	.19	.77	-.40
Cooperativeness	.06	.30	.38	-.47
Conscientiousness	-.17	.31	.31	-.38
Nondelinquency	.13	.31	.63	-.42
Traditional Values	-.24	.25	1.00	-.40
Work Orientation	-.33	.30	.32	-.38
Internal Control	.03	.15	.22	-.44
Energy Level	-.12	.24	.45	-.41
Dominance (Leadership)	-.63	.25	.32	-.38
Physical Condition	-.07	.20	.35	-.39

[a] Standard mean differences are [Mean (Honest) minus Mean (Fake)]/SD (Honest).

[b] Correlations are average of correlations for first administration under Honest and relevant Fake condition.

formulas developed on these data cross validate and do they increase criterion-related validity?

Overall, the ABLE data from the Fort Bragg faking study show that:

1. Soldiers can distort their responses when instructed to do so.

2. The response validity scales detect intentional faking; Unlikely Virtues (Social Desirability) detects Faking Good and Non-Random Response, Poor Impression, and Self-Knowledge detect Faking Bad.

3. An individual's Unlikely Virtues (Social Desirability) scale score can be used to adjust his or her content scale scores to reduce variance associated with faking good; an individual's Poor Impression scale score can be used to adjust his or her content scale scores to reduce variance associated with faking bad.

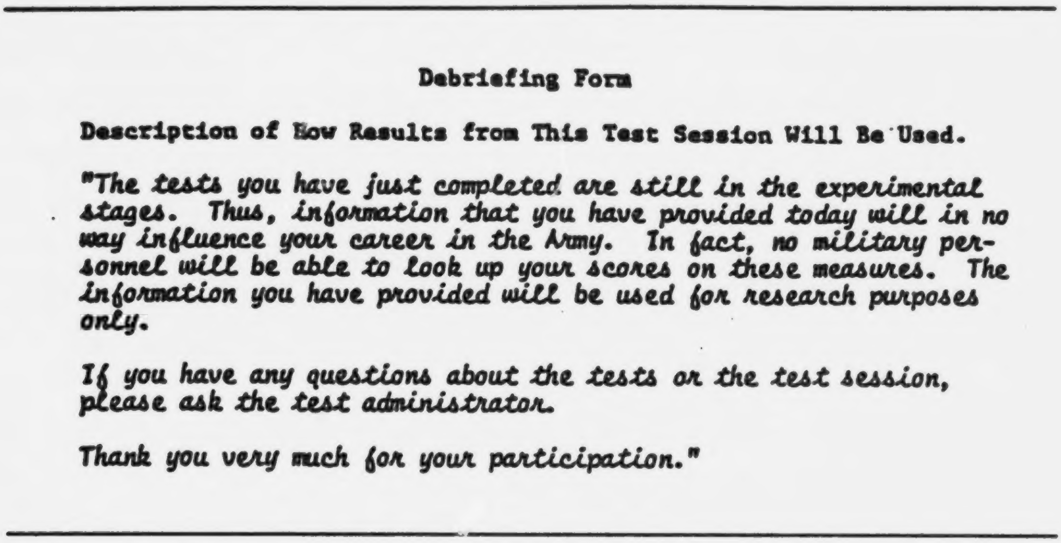
Faking in An Applicant Setting

MEPS "Applicant" Sample. Another of the purposes of the fakability study was to determine the extent to which intentional distortion actually is a problem in an applicant setting. To investigate this question, the ABLE and AVOICE were administered at the Minneapolis MEPS. However, the sample of 126 recruits who completed the inventories were not true "applicants," in that they had just recently been sworn into the Army.

MEPS Procedures. To approximate the applicant response set as closely as was possible with this sample, recruits were allowed to believe that their scores on these inventories might affect their Army careers. This was accomplished by deleting all references in the standard Privacy Act Statement (given to all subjects at the beginning of a testing session) to these data being collected for research purposes only, and not having any effect on the participant's career or status in the Army. Recruits were then asked to complete the ABLE and AVOICE, after which they were debriefed. In the debriefing each recruit was asked to read the debriefing form displayed as Figure 8.3, and the administrator orally summarized the information on this form and answered any questions the recruit might have.

To examine the extent to which recruits actually believed their ABLE and AVOICE scores would have an effect on their Army career, each recruit filled out the single-item form shown in Figure 8.4 prior to debriefing. Of the 126 recruits in this sample, 57 responded "yes" to this question, 61 said "no," and 8 wrote in that they didn't know. Thus, while the MEPS sample is not a true "applicant" sample, its make-up (recently sworn-in recruits, close to half of whom believe their ABLE and AVOICE scores will affect their Army career) is reasonably close. The response set for this sample is almost certainly more similar to that of the applicant population than is the Fort Knox sample.

MEPS Results Compared With Fort Knox and Fort Bragg Data. Table 8.15 shows mean scores for MEPS recruits and the two "Honest" conditions of this study at Fort Bragg and Fort Knox. Even though the recruits are probably trying not to create a poor impression (MEPS Poor Impression mean is 1.05,



Debriefing Form

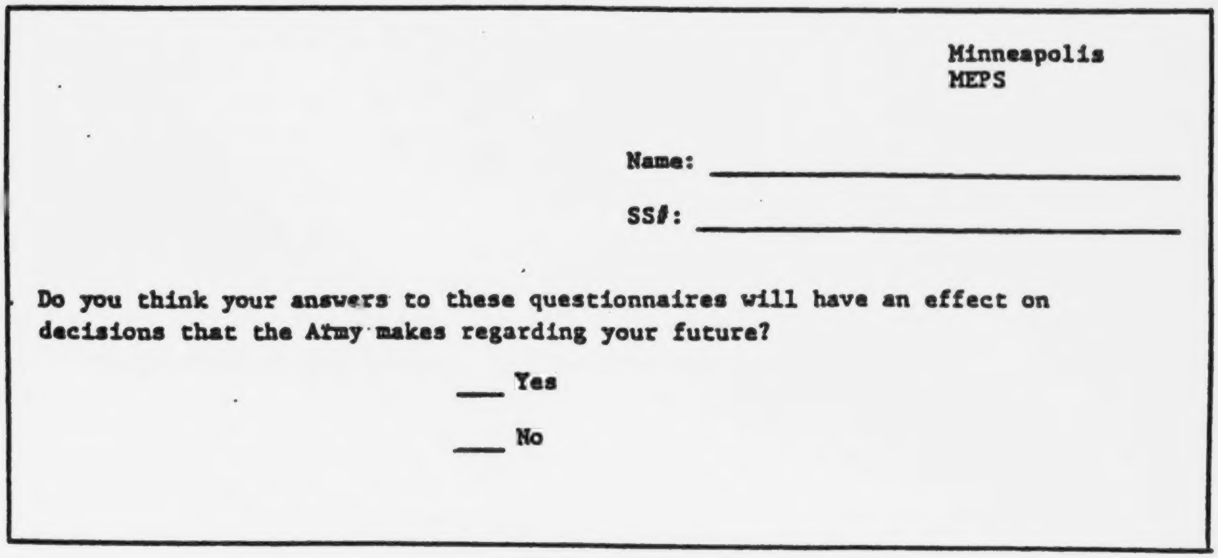
Description of How Results from This Test Session Will Be Used.

"The tests you have just completed are still in the experimental stages. Thus, information that you have provided today will in no way influence your career in the Army. In fact, no military personnel will be able to look up your scores on these measures. The information you have provided will be used for research purposes only.

If you have any questions about the tests or the test session, please ask the test administrator.

Thank you very much for your participation."

Figure 8.3 Debriefing Form used in the faking study at the Military Entrance Processing Station (MEPS).



Minneapolis
MEPS

Name: ______________________

SS#: ______________________

Do you think your answers to these questionnaires will have an effect on decisions that the Army makes regarding your future?

___ **Yes**

___ **No**

Figure 8.4 Form filled out by MEPS recruits before debriefing.

Table 8.15

Comparison of Results From Fort Bragg Honest, Fort Knox, and MEPS (Recruits) ABLE Scales

ABLE Scale	Fort Bragg Honest[a] N	Fort Bragg Honest[a] Mean	MEPS (Recruits) N	MEPS (Recruits) Mean	Fort Knox N	Fort Knox Mean	Total SD	Degrees of Freedom	F	p
Response Validity Scales										
Social Desirability (Unlikely Virtues)	116	15.91	121	16.63	276	16.60	3.21	2,510	2.15	.12
Self-Knowledge	116	29.54	121	28.03	276	29.64	3.63	2,510	9.10	.00
Non-Random Response	116	7.58	121	7.79	276	7.75	.64	2,510	3.73	.02
Poor Impression	116	1.50	121	1.05	276	1.54	1.84	2,510	3.15	.04
Content Scales										
Emotional Stability	112	66.22	118	66.03	272	65.05	7.86	2,499	1.18	.31
Self-Esteem	112	34.77	118	34.04	272	35.12	5.00	2,499	1.93	.15
Cooperativeness	112	53.33	118	54.60	272	54.19	6.05	2,499	1.34	.26
Conscientiousness	112	46.37	118	46.49	272	48.97	5.86	2,499	12.24	.00
Nondelinquency	112	53.24	118	54.36	272	55.49	6.91	2,499	4.48	.01
Traditional Values	112	36.67	118	36.97	272	37.28	4.50	2,499	.77	.46
Work Orientation	112	59.71	118	58.37	272	61.40	7.73	2,499	6.90	.00
Internal Control	112	49.48	118	51.90	272	50.37	6.13	2,499	4.75	.01
Energy Level	112	57.56	118	56.67	272	57.19	6.95	2,499	.48	.62
Dominance (Leadership)	112	35.54	118	32.84	272	35.41	6.05	2,499	8.69	.00
Physical Condition	112	32.96	118	28.27	272	31.08	7.49	2,499	12.10	.00

---

[a] Scores are based on persns who responded to the Honest condition first.

which is lower than both the Fort Knox and Fort Bragg means, 1.54 and 1.50, respectively), they do not score significantly higher on the response validity scale Unlikely Virtues (Social Desirability). Indeed their mean score is lowest on six of the 11 content scales, scales on which it would be desirable to score high rather than low. They score highest on only two content scales and only one, Internal Control, is significant.

In sum, intentional distortion may not be a significant problem in an applicant setting. (What faking or distortion would be in a draft situation cannot be estimated in the present non-draft situation in the United States).

Faking Study Results - Interests Inventory

We divided the interest scales into the two groups, combat-related and combat support, that emerged when we factor analyzed the AVOICE Fort Knox data. We then performed a multivariate analysis of variance (MANOVA) on the experimental data from Fort Bragg. Tables 8.16 and 8.17 show the findings for the interactions, the sources of variance most relevant to the question of whether soldiers can or cannot intentionally distort their responses.

As can be seen, 9 of the 11 combat-related AVOICE scales are sensitive to intentional distortion, and 9 of the 12 combat support scales are sensitive to intentional distortion. The interaction of Fake x Set x Order is either significant or marginally significant, indicating that order of conditions in which the participants completed the AVOICE also affected the result.

Tables 8.18 and 8.19 show mean scores for the various conditions when the particular condition was the first administration. When told to distort their responses so that they would not be likely to be placed in combat-related occupational specialties (MOS), that is, instructed to Fake Noncombat, soldiers tended to decrease their scores on all scales. Scores on 19 of 24 interest scales were lower in Fake Noncombat as compared to the honest condition. In the Fake Combat condition, soldiers in general increased their combat-related scale scores and decreased their combat support scale scores.

We next examined the extent to which the ABLE response validity scales, which had demonstrated they could detect intentional distortion, could be used to adjust AVOICE scale scores for faking combat and faking noncombat. Table 8.20 shows the adjusted mean differences in AVOICE scale scores after regressing out ABLE Social Desirability and Poor Impression. Comparing these differences to the unadjusted differences shown in Tables 8.18 and 8.19 reveals that these adjustments have little effect, perhaps because the response validity scales consisted of items from the ABLE and the faking instructions for the ABLE and AVOICE were different. The ABLE faking instructions were Fake Good and Fake Bad, whereas the AVOICE faking instructions were Fake Combat and Fake Noncombat.

As in the ABLE, the question was investigated of whether or not applicants would, in fact, tend to distort their responses to the AVOICE. Tables 8.21 and 8.22 show the mean scores for the MEPS recruits and the two Honest conditions, Fort Bragg and Fort Knox. There appears to be no parti-

Table 8.16

Fakability Study, MANOVA Results for AVOICE Combat-Related Scales: Fort Bragg (N = 164)

Type and Name of Scale	Interactions Fake x Set	 Fake x Set x Order
Combat-Related Scales		
Overall	S	NS*
Marksman	S	---
Agriculture	S	---
Armor/Cannon	S	---
Vehicle/Equipment Operator	S	---
Outdoors	S	---
Infantry	S	---
Law Enforcement	S	---
Heavy Construction/Combat	S	---
Mechanics	NS	---
Electronics	NS	---
Adventure	S	---

Note: S = Significant, p≤.01.
NS = Nonsignificant, p>.01.
* = Marginally significant, .05≤p>.01.

Table 8.17

Fakability Study, MANOVA Results for AVOICE Combat Support Scales: Fort Bragg (N = 201)

	Interactions	
Type and Name of Scale	Fake x Set	Fake x Set x Order
Combat Support Scales		
Overall	S	S
Mathematics	S	NS
Aesthetics	S	S
Leadership	S	S
Electronic Communication	S	S
Automated Data Processing	S	S
Teaching/Counseling	NS	NS
Drafting	NS	NS
Audiographics	NS	NS
Science/Chemical Operations	S	NS
Supply Administration	S	NS
Office Administration	S	NS
Medical Services	NS*	NS
Food Service	S	NS*

Note: S = Significant, $p \leq .01$.
NS = Nonsignificant, $p > .01$.
* = Marginally significant, $.05 \leq p > .01$.

Table 8.18

Effects of Faking, AVOICE Combat Scales: Fort Bragg

	Honest[a]			Fake Combat[a]			Fake Noncombat[a]			Effect Size[b]	
AVOICE Combat Scales	N	Mean	SD	N	Mean	SD	N	Mean	SD	Honest vs. Combat	Honest vs. Noncombat
Marksman	122	18.1	4.5	58	20.2	3.9	60	12.8	5.9	-.49	1.06
Agriculture	124	15.0	3.8	59	12.9	3.6	60	15.1	4.0	.56	-.03
Armor/Cannon	124	24.2	5.8	59	28.9	7.6	60	15.1	6.3	-.73	1.53
Vehicle/Equipment	124	28.7	6.4	59	26.6	7.9	60	23.5	8.0	.30	.75
Outdoors	123	36.0	6.1	59	38.3	6.0	60	25.7	10.2	-.38	1.34
Infantry	123	33.5	6.8	59	37.8	8.2	59	20.5	8.4	-.59	1.77
Law Enforcement	124	53.3	10.8	59	54.5	12.1	60	42.3	12.5	-.11	.97
Heavy Construction	124	70.5	16.3	59	68.9	15.0	59	58.7	16.4	.10	.72
Mechanics	124	50.7	12.7	59	44.6	15.2	60	47.3	13.6	.45	.26
Electronics	124	58.1	18.3	59	50.3	17.3	60	56.8	18.0	.43	.07
Adventure	108	37.5	4.3	56	38.1	3.7	54	26.8	6.6	-.15	2.06

[a] Values are based on the sample that completed the questionnaire under the condition of interest first.

[b] Effect Size = (Mean Honest minus Mean Combat, or Noncombat)/SD Total

Table 8.19

Effects of Faking, AVOICE Combat Support Scales: Fort Bragg

AVOICE Combat Support Scales	Honest[a]			Fake Combat[a]			Fake Noncombat[a]			Effect Size[b]	
	N	Mean	SD	N	Mean	SD	N	Mean	SD	Honest vs. Combat	Honest vs. Noncombat
Mathematics	120	14.2	4.7	56	11.8	4.7	59	15.6	5.0	.51	-.29
Aesthetics	120	14.6	4.1	57	12.1	4.6	59	17.1	5.5	.59	-.54
Leadership	124	22.3	4.2	59	21.5	4.1	59	17.3	5.8	.19	1.05
Electronic Communication	123	21.1	6.1	59	21.8	7.0	60	14.2	5.6	-.11	1.16
Automated Data Processing	122	20.4	6.7	58	15.5	7.2	59	23.8	7.4	.71	-.49
Teaching/Counseling	124	23.8	5.7	59	20.7	5.6	60	21.0	5.6	.55	.49
Drafting	124	22.3	6.1	59	18.4	6.2	60	21.5	5.5	.64	.14
Audiographics	124	23.5	5.6	59	18.7	6.2	60	20.7	5.6	.83	.50
Science/Chemical Operations	123	28.0	8.4	59	28.0	9.2	60	25.8	9.6	0	.25
Supply Administration	124	30.5	9.8	59	26.4	9.6	60	35.3	11.9	.42	-.46
Office Administration	123	38.5	13.3	59	31.2	12.3	59	49.5	17.2	.56	-.75
Medical Services	124	67.8	18.3	59	60.4	17.5	60	61.0	17.8	.41	.37
Food Service	122	38.0	10.2	59	31.0	10.6	59	45.8	16.3	.68	-.62

---

[a] Values are based on the sample that completed the questionnaire under the condition of interest first.

[b] Effect Size = (Mean Honest minus Mean Fake Combat, or Fake Noncombat)/SD Total

Table 8.20

Effects of Regressing Out Response Validity Scales (Unlikely Virtues and Poor Impression) on Faking Condition, AVOICE Combat Scales Scores: Fort Bragg

	Fake Combat		Fake Noncombat	
	Adjusted Standardized Mean Difference[a]	Correlation with Social Desirability[b]	Adjusted Standardized Mean Difference[a]	Correlation with Poor Impression[b]
Combat-Related AVOICE Scales				
Marksman	-.71	.08	1.31	-.14
Agriculture	.48	-.15	-.14	.11
Armor/Cannon	-1.35	.19	1.08	-.15
Vehicle/Equipment Operator	-.39	-.02	.59	.01
Outdoors	-.33	.03	1.82	-.27
Infantry	-1.08	.08	1.38	-.18
Law Enforcement	-.12	-.03	.86	-.13
Heavy Construction/Combat	-.06	.02	.32	-.15
Mechanics	.32	.02	.47	-.04
Electronics	-.03	-.02	.30	-.08
Adventure (ABLE)	.09	-.02	-1.09	-.45

(Continued)

Table 8. 20 (Continued)

Effects of Regressing Out Response Validity Scales (Unlikely Virtues and Poor Impression) on Faking Condition, AVOICE Combat Scales Scores: Fort Bragg

	Fake Combat		Fake Noncombat	
Combat Support AVOICE Scales	Adjusted Standardized Mean Difference[a]	Correlation with Social Desirability[b]	Adjusted Standardized Mean Difference[a]	Correlation with Poor Impression[b]
Mathematics	.23	-.05	-.34	.22
Aesthetics	.51	-.24	-.17	.13
Leadership	-.27	-.05	1.28	-.16
Electronic Communication	-.55	.15	.78	-.26
Automated Data Processing	.76	-.06	-.29	.03
Teaching/Counseling	.56	-.13	.76	.03
Drafting	.26	-.05	.48	.14
Audiographics	.69	-.08	.37	.05
Science/Chemical Operations	-.53	.01	.10	.04
Supply Administration	.05	-.05	.04	.21
Office Administration	.37	-.10	-.15	.24
Medical Services	.00	-.11	.33	.10
Food Service	.39	-.13	-.75	.26

---

[a] Standardized mean differences are [mean (Honest) - mean (Fake)]/SD (Honest).

[b] Correlations are average of correlations for first administration under honest and relevant fake condition.

Table 8.21

Comparison of Fort Bragg Honest, Fort Knox, and MEPS (Recruits) AVOICE Combat-Related Scales[a]

Combat-Related AVOICE Scales	Fort Bragg[b] (Honest) N	Fort Bragg[b] (Honest) Mean	MEPS (Recruits) N	MEPS (Recruits) Mean	Fort Knox N	Fort Knox Mean	Pooled SD	Degrees of Freedom	F	P
Marksman	122	18.1	121	17.0	256	15.8	4.4	2,496	12.0	.00
Agriculture	124	15.0	124	15.4	267	14.1	3.7	2,512	4.5	.01
Armor/Cannon	124	24.2	125	27.0	268	22.4	6.2	2,514	22.8	.00
Vehicle/Equipment Operator	124	28.7	125	31.0	268	28.1	7.2	2,514	7.3	.00
Outdoors	123	36.0	125	35.2	268	31.7	6.1	2,513	26.7	.00
Infantry	123	33.5	125	33.2	268	29.1	6.8	2,513	24.8	.00
Law Enforcement	124	53.3	124	48.4	265	48.1	11.3	2,510	10.1	.00
Heavy Construction/Combat	124	70.5	124	70.6	269	65.8	16.8	2,514	5.3	.01
Mechanics	124	50.7	125	53.4	269	50.0	14.1	2,515	2.54	.08
Electronics	124	58.1	125	59.5	266	60.0	17.5	2,512	0.5	.62
Adventure (ABLE)	108	37.5	101	35.5	211	32.8	5.2	2,417	31.5	.00

---

[a] MANOVA significant using Wilk's Lambda (F = 6.3; df = 22,754; p = .00)

[b] Fort Bragg data are for honest first condition only.

Table 8.22

Comparison of Fort Bragg Honest, Fort Knox, and MEPS (Recruits) AVOICE Noncombat-Related Scales[a]

Noncombat-related AVOICE Scales	Fort Bragg[b] (Honest)		MEPS (Recruits)		Fort Knox		Pooled	Degrees of		
	N	Mean	N	Mean	N	Mean	SD	Freedom	F	p
Mathematics	120	14.2	122	13.7	252	15.1	4.4	2,491	4.7	.01
Aesthetics	120	14.7	121	13.8	261	14.3	4.2	2,499	4.2	.02
Leadership	124	22.3	125	19.7	269	20.3	4.5	2,515	11.9	.00
Electronic Communication	123	21.1	125	21.7	268	21.1	5.7	2,513	0.4	.67
Automated Data Processing	122	20.4	121	19.0	256	23.3	6.3	2,496	22.1	.00
Teaching/Counseling	124	23.8	125	21.0	268	22.9	8.7	2,514	8.7	.00
Drafting	124	22.3	125	20.7	270	21.5	6.1	2,516	1.6	.21
Audiographics	124	23.5	124	22.1	269	23.8	5.6	2,514	4.2	.02
Science/Chemical Operations	123	28.0	125	26.9	269	29.3	8.8	2,514	3.5	.03
Supply Administration	124	30.5	125	33.1	268	34.6	9.8	2,515	8.9	.00
Office Administration	123	38.5	125	38.0	267	45.2	12.7	2,512	19.4	.00
Medical Services	124	67.8	125	61.1	267	68.5	18.8	2,513	6.9	.00
Food Service	122	38.0	125	42.4	269	42.2	10.8	2,513	7.4	.00

---

[a] MANOVA significant using Wilk's Lambda (F = 6.1; df = 26,896; p = .00).

[b] Fort Bragg data are for honest first condition only.

cular pattern to the mean score differences. The applicants score lowest, highest, and in the middle about an equal number of times.

Overall, the AVOICE data from the faking study show that:

1. Soldiers can distort their responses when instructed to do so.

2. The ABLE Social Desirability and Poor Impression scales are not as effective for adjusting AVOICE scale scores in the faking conditions of Combat/Noncombat as they are for adjusting ABLE content scale scores in the Faking Good/Faking Bad conditions.

3. Faking or distortion may not be a significant problem in an applicant setting.

CONCLUDING COMMENTS

The field tests of the non-cognitive measures indicate they are good measures of the intended constructs and that they are likely to contribute unique, reliable variance to the predictor domain. Score distributions and reliabilities show the measures to be sound psychometrically. The uniqueness analyses showed that the ABLE and AVOICE scales are measuring individual differences largely independent from those measured via the ASVAB or other parts of the Pilot Trial Battery. Factor analyses of ABLE and AVOICE scales showed a relatively simple underlying structure that makes intuitive sense. Investigations of faking and fakability indicate scores can be intentionally distorted when persons are instructed to do so, but distortion does not appear to occur in the present applicant setting, and the response validity scales on the ABLE can probably be used to correct for distortion when it does occur. However, more research is needed on the methods of applying such corrections and the effects of such corrections on the validity of the non-cognitive scales for predicting job performance or other important criteria.

Chapter 8 Reference

Hough, L. M., Barge, B. N., Houston, J. S., McGue, M. K., & Kamp, J. D. (1985). *Problems, issues and results in the development of temperament, biographical, and interest measures.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

CHAPTER 9

FORMULATION OF THE TRIAL BATTERY

**Norman G. Peterson, Jeffrey J. McHenry, Marvin D. Dunnette,
Jody L. Toquam, Leaetta M. Hough, Bruce N. Barge, Rodney L. Rosse,
Janis S. Houston, and VyVy A. Corpe**

The way in which the Pilot Trial Battery was revised to produce the Trial Battery is described in this chapter. The previous chapters have presented and discussed the development, pilot tests, and field tests of the Pilot Trial Battery. They show, we think, that the Pilot Trial Battery measures, as a whole, are psychometrically sound, measure relatively unique constructs, and appear to hold considerable promise as predictors of various important criteria of job performance for Army soldiers. The nature of the revisions described here focused on satisfying the pragmatic criteria of limited testing time available for future Project A research, as well as improving the measures in the Pilot Trial Battery.

REVISIONS TO THE PILOT TRIAL BATTERY

The full Pilot Trial Battery, as administered at the field tests, required approximately 6.5 hours of actual administration time. However, the Trial Battery developed from the Pilot Trial Battery (see Figure 1.2) had to be administered in about 4 hours during the next phase of the project (Concurrent Validation). Therefore, not only did the measures in the Pilot Trial Battery need revision on the basis of field test experience, but the total length of the battery had to be reduced by 33 percent.

We devised three general principles, which we called a strategy, to be used as a guide in making the revision and reduction decisions. These principles were consonant with the theoretical and practical orientation that had been used since the inception of the project, as described in Chapter 1. The principles were:

- Maximize the heterogeneity of the battery by retaining measures of as many different constructs as possible.
- Maximize the chances of incremental validity and classification efficiency as much as possible.
- Retain measures with adequate reliability.

Five more concrete implications or guidelines for adopting this strategy were developed. These are shown in Figure 9.1. With these guidelines in mind, Task 2 staff prepared summaries and presentations of the information described in Chapters 2 through 8.

In March 1985, these presentations were made at an In Progress Review (IPR) meeting held to consider the field test data and other relevant information, and decide on the methods and nature of revising the Pilot Trial Battery. Generally speaking, the presentations were within the three domains--cognitive (paper and pencil), perceptual/psychomotor (computer-administered), and non-cognitive-- that had been used throughout the research (point 1 in Figure 9.1). The psychometric characteristics of each measure within a domain were reported, followed by a presentation of the covariance (correlations and factor structure) of the measures within the domain, across the domains, and with the ASVAB (uniqueness analyses). Then, estimates of expected validities for training and job performance criteria (based on the expert judgments, literature review, and Preliminary Battery analyses) were presented. Finally, initial recommendations for reduction and revisions were made.

Considerable discussion was generated by these presentations, but the IPR group reached a consensus on the reductions and revisions to be made to the Pilot Trial Battery. This set of recommendations was the presented to and discussed at the meeting of the full Scientific Advisory Group. A few changes were made at this meeting.

---

1. Retain Measures in All Three Predictor Areas:
    - Cognitive (Paper-and-Pencil)
    - Perceptual/Psychomotor (Computer-Administered)
    - Non-Cognitive (Paper-and-Pencil)
2. Retain Measures That Add Unique Variance
    - Variance Not Accounted for by ASVAB
    - Variance Not Accounted for by Other Pilot Trial Battery Measures
3. Retain Measures That Predict Training Success and/or for which Experts or Literature Review Suggests Validity for Job Performance, Especially for Important Criteria or Criteria Not Presently Predicted by ASVAB
4. Retain Measures That Show Stability With Respect to:
    - Test-Retest
    - Practice
    - Faking/Fakability
5. Within Measures, Retain Items That Measure the Dominant Construct and Maximize Content Coverage

---

Figure 9.1 Guidelines for evaluating and retaining Pilot Trial Battery measures in order to produce the Trial Battery.

Tables 9.1, 9.2, and 9.3 summarize the change recommendations that came from these meetings. These recommendations were used to guide the development of the Trial Battery from the Pilot Trial Battery. In the following sections, we describe these changes and their rationales, plus any internal improvements made to each measure.

Changes to Cognitive Paper-and-Pencil Tests

Analyses of pilot and field tests of the cognitive paper-and-pencil tests showed that the tests, as a group, measure various aspects of spatial ability. When factor-analyzed with ASVAB subtests and the computer-administered tests from the Pilot Trial Battery, they formed a single factor of their own (see Table 6.10). Factor analysis of the tests by themselves, however, tends to show four or five factors (see Table 3.13). These results are not surprising, but we point them out to illustrate the point that the identification of the number and type of constructs measured by a set of tests depends very much on the level of analysis a researcher chooses. For purposes defined here, that is, reducing the number of tests to carry forward from the Pilot Trial Battery to the Trial Battery, we focused on a more specific level (four-five factors), but kept in mind that all the tests measure an underlying, more global spatial ability. Changes to the cognitive tests for use in the Trial Battery are described in the context of the constructs the tests were designed to measure: Spatial Visualization-Rotation and Field Independence, Spatial Visualization-Scanning, Figural Reasoning/Induction, and Spatial Orientation.

In the Pilot Trial Battery, the Spatial Visualization--Rotation and Field Independence construct was measured by three tests: Assembling Objects, Object Rotation, and Shapes. Although Shapes was originally designed to measure Field Independence, and pilot test results indicated it correlated .50 with a marker test of that ability, we considered this test in concert with the two Rotation tests for purposes of reducing the size of the Pilot Trial Battery. This combination seemed justified because the three tests had a similar pattern of factor loadings (see Table 3.13). The Shapes Test was dropped because the evidence of validity for job performance for tests of this type was judged to be less impressive than for the other two tests. The Object Rotation Test was not changed. Eight items were dropped from the Assembling Objects Test by eliminating those items that were very difficult or very easy, or had low item-total correlations. The time limit for Assembling Objects was not changed. The effect was to make Assembling Objects more a power test than it was prior to the changes.

The Spatial Visualization-Scanning construct was measured by two Pilot Trial Battery tests, Mazes and Path. The Path Test was dropped and the Mazes Test was retained with no changes. Mazes showed higher test-retest reliabilities than Path (.71 vs .64) and lower gain scores (.24 SD units for Mazes vs .62 SD units for Path), which was desirable. In addition, Mazes was a shorter test than Path (5.5 minutes vs 8 minutes).

The Figural Reasoning/Induction construct was measured by the Reasoning 1 and Reasoning 2 tests. Reasoning 1 was evaluated as the better of the two tests because it had higher reliabilities for both internal consis-

Table 9.1

Summary of Changes to Cognitive Paper-and-Pencil Measures in the Pilot Trial Battery

Test Name	Changes
Assembling Objects	Decrease from 40 to 32 items.
Object Rotation	Retain as is with 90 items.
Shapes	Drop Test.
Mazes	Retain as is with 24 items.
Path	Drop Test.
Reasoning 1	Retain as is with 30 items. New name REASONING TEST.
Reasoning 2	Drop Test.
Orientation 1	Drop Test.
Orientation 2	Retain as is with 24 items. New name ORIENTATION TEST.
Orientation 3	Retain as is with 20 items New name MAP TEST

Table 9.2

Summary of Changes to Computer-Administered Pilot Trial Battery Measures

Test Name	Changes
COGNITIVE/PERCEPTUAL TESTS	
Demographics	Eliminate race, age, and typing experience items. Retain SSN and video experience items.
Simple Reaction Time	No changes.
Choice Reaction Time	Increase number of items from 15 to 30.
Perceptual Speed & Accuracy	Reduce items from 48 to 36. Eliminate word items.
Target Identification	Reduce items from 48 to 36. Eliminate moving items. Allow stimuli to appear at more angles of rotation.
Short-Term Memory	Reduce items from 48 to 36. Establish a single item presentation and probe delay period.
Cannon Shoot	Reduce items from 48 to 36.
Number Memory	Reduce items from 27 to 18. Shorten item strings. Eliminate item part delay periods.
PSYCHOMOTOR TESTS	
Target Tracking 1	Reduce items from 27 to 18. Increase item difficulty.
Target Tracking 2	Reduce items from 27 to 18. Increase item difficulty.
Target Shoot	Reduce items from 40 to 30 by eliminating the extremely easy and extremely difficult items.

Table 9.3

Summary of Changes to Pilot Trial Battery Versions of Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE)[a]

Inventory/Scale Name	Changes
ABLE Total	Decrease from 270 to approximately 199 items.
AVOICE Total	Decrease from 309 to approximately 228 items.
AVOICE Expressed Interest Scale	Drop scale.
AVOICE Single Item Holland Scales	Drop scales.
AVOICE Agriculture Scale	Drop scale.
Organizational Climate/Environment Preference Scales	Move to criterion measure booklet (delete from AVOICE booklet).

[a] In addition to the changes outlined in this table by inventory/scale, it was recommended that all ABLE item response options be standardized as three-option responses and all AVOICE item response options be standardized as five-option responses.

tency (alpha = .83 vs .65 and separately timed, split-half coefficients = .78 vs .63) and test-retest (.64 vs .57), as well as a higher uniqueness estimate (.49 vs .37). Reasoning 1 was retained with no item or time limit changes and Reasoning 2 was dropped. Reasoning 1 was renamed Reasoning Test.

Three tests measured the Spatial Orientation construct in the Pilot Trial Battery. Orientation 1 was dropped because it showed lower test-retest reliabilities (.67 vs .80 and .84) and higher gain scores (.63 SD units vs .11 and .08 SD units). In addition, we modified the instructions for Orientation 2 because field test experience had indicated that the PTB instructions were not as clear as they should be. Orientation 2 was renamed Orientation Test. Orientation 3 was retained with no changes and renamed Map Test.

Changes to Perceptual/Psychomotor Computer-Administered Tests

Before describing the changes made to specific perceptual/psychomotor tests in the computer-administered battery, we describe several improvements to the computer battery as a whole.

Modifications in Computer Administration Procedures. The general changes included the following:

1. Virtually all test instructions were modified, in these ways:

   - Most instructions were shortened considerably.

   - Names of buttons, slides, and switches on the response pedestals were written in capital letters whenever they appeared in the instructions (e.g., BLUE, VERTICAL, RIGHT) to attract subjects' attention faster and more effectively.

   - Test terms and jargon were standardized. For example, in the PTB test instructions, the response pedestal was at various times called the "testing panel," the "response panel," and the "response pedestal." In the Trial Battery instructions, this apparatus was always referred to as the "response pedestal."

   - Where possible, the following standard outline was used in preparing the instructions:

     -- Test name
     -- One-sentence description of the purpose of the test
     -- Step-by-step test instructions
     -- One practice item
     -- Brief re-statement of test instructions
     -- Two or three additional practice items
     -- Instructions to call test administrator if there are questions about the test

2. Whenever test items had a correct response, the subject was given feedback on the practice items to indicate whether he/she had answered the item correctly.

3. Rest periods were eliminated from the battery. (Previously, there were rest periods between the first half and second half of the items within several of the tests.) This was feasible because most tests were shortened.

4. The computer programs controlling test administration were merged into one super-program, eliminating the time required to load the programs between tests.

5. The format and parameters used in the software containing test items were reworded, so that the software was more "self-documented."

6. The total time allowed for subjects to respond to a test item (or, in other words, response time limit) was set at 9 seconds for all reaction time tests (Simple and Choice Reaction Time, Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification). In the PTB version the response time limit had varied from test to test, for no particular reason. The field test data showed that, on almost all trials of all reaction time tests, subjects were able to respond within 9 seconds. Therefore, the 9-second time limit was adopted as a standard.

7. Also, with regard to the reaction time tests, the software was changed so that the stimulus for an item disappeared when the subject lifted his/her hand from the home button (in order to make a response). Subjects are instructed not to lift their hands from the home buttons until they have determined the correct response; in this manner, separate measures of decision and movement time can be obtained. However, more than a few of the field test subjects continued to study the item stimulus to determine the correct response after leaving the home buttons. By causing the item to disappear, we hoped to eliminate that problem.

All of the changes to the overall computer-administered test battery described above, and the individual test changes described below, were subjected to a series of small sample tryouts (N < 6 in each tryout) at the Minneapolis Military Entrance Processing Station (MEPS). These tryouts were for the purpose of inspecting and evaluating the software changes (including test items), eliciting feedback about instruction changes and insuring that the time needed to take the computer-administered test battery was within the time that would be available for the upcoming Concurrent Validation phase of Project A. No data were analyzed as a result of these tryouts because the total N was too small (less than 40), but they fulfilled the purpose of insuring that all changes were made correctly and were achieving the end desired.

Changes to Content of Tests Administered by Computer. We turn now to a description of the specific changes made to the individual computer-administered tests for use in the Trial Battery.

In the demographic section of the computer battery, items asking about age, race, and typing experience were deleted. Information on age and race is available from other sources. Typing experience is no longer relevant since subjects' responses are now obtained via the response pedestal instead of a standard keyboard.

No changes were recommended for Simple Reaction Time. However, we re-randomized the order of the pretrial intervals (the interval between the time the subject depresses the home button keys and the appearance of the trial stimulus). This was done because the pretrial intervals (the order of these intervals had been randomly determined) tended to increase over trials 7-14, then dropped precipitously for trial 15; as a result, mean response time for trial 15 was significantly higher than mean response times for the previous several trials. Re-randomization was therefore considered desirable, to remove this abnormality.

The number of items in Choice Reaction Time was increased from 15 to 30 in an attempt to increase the test-retest reliability for mean reaction time on this test.

Twelve items were eliminated from Perceptual Speed and Accuracy (reduced from 48 to 36 items), primarily to save time. Internal consistency estimates were high for scores on this test (.83, .96, .88, and .74 for Percent Correct, Mean Reaction Time, Slope, and Intercept, respectively), so item reduction did not seem to be cause for concern in that regard. Test-retest reliabilities were lower than internal consistencies, but it was not clear that item reduction would affect this greatly. The 12 items eliminated were all the "word" items (see Chapter 5 for a description of the item types in this test) rather than any of the alpha, numeric, symbolic, or mixed items, because word items were not used to calculate two of the scores, Slope and Intercept.

Several changes were made to the Target Identification Test. First, one of the two item types--the "moving" items--was eliminated. Field test data showed that scores on the "moving" and stationary items correlated .78, and the moving items had lower test-retest reliabilities than stationary items (.54 vs .74) and also had lower uniqueness estimates (.44 vs .56). Also, two item parameters were modified. All target objects were made the same size (50% of the size of the objects depicted as possible answers) since field test analyses indicated size had had no appreciable effect on reaction time. A third level of angular rotation was added so the target objects were rotated either 0°, 45°, or 75°. Theoretically, and as found in past research, reaction time is expected to increase with greater angular rotation. Two of the item parameters were not changed (position of correct response object and direction of target object). Finally, the number of items was reduced from 48 to 36 in order to save time. Internal consistency and test-retest estimates indicated that the level of risk attached to this reduction would be acceptable. (For Mean Reaction Time, the internal consistency estimate was .96 and the test-retest estimate was .67.) The reduction from 48 to 36 items was accomplished by retaining 12 of the 24 "moving" items (which were to be eliminated as an item type, see just above) as stationary items. That is, the items had the same parameters they possessed as "moving" items, but were presented as "stationary" items. The retained items were those that had the proper item parameters to allow a balanced number of items in each of the cells defined by crossing the item parameters. The test, as modified, had two items in each of 18 cells determined by crossing angular rotation (0°,45°, 75°), position of correct response object (left, center, or middle of screen), and direction of target object (left-facing or right-facing).

One item parameter, probe delay period, was eliminated from the Short-Term Memory Test, while two others, item type (symbolic vs. letter) and item length (1, 2, or 5 objects) were retained. Analyses of field test data showed that probe delay period did not significantly affect Mean Reaction Time scores. To save time, 12 items were eliminated. (Eliminating the probe delay period did not result in any reduction in items.) Two of the three most important scores for this test appeared to have high enough reliabilities to withstand such a reduction (internal consistency and test-retest estimates were .94 and .78, respectively for Mean Reaction Time, .52 and .47 for Slope, and .84 and .74 for Intercept). Items were eliminated by deleting those items that had the lowest item-total score correlation, within the limitation of maintaining balance in the distribution of items across the cells defined by item parameters.

Finally, the software controlling the administration of this test was rewritten in an attempt to reduce the amount of missing data occurring on the test. Field test data indicated that some subjects apparently did not completely understand the instructions, and completed items inappropriately, causing missing data (specifically, they released the home buttons after the item's stimulus set disappeared but before the probe appeared). The rewritten software gave feedback to the subject if an item was inappropriately completed. If a subject completed three items inappropriately, he/she was told (by a message on the screen) to call the test administrator for further instruction; also, the test would not continue until the administrator made a sequence of button pushes (unknown to the subject).

The number of items on the Cannon Shoot Test was reduced from 48 to 36, again to save time. Internal consistency and test-retest reliabilities for the Time Error Score were high enough (.88 and .66, respectively) to warrant such reduction without the expectation of a significant impact on reliability.

Also, the items were modified to eliminate two problems observed during the field tests. First, on some items, the target was actually not on the screen as it began its movement toward the cannon's line of fire. Second, on some items, the subject had to fire at the target almost as soon as it appeared on the screen in order to hit the target with the cannon shell. Such items provided subjects with little or no opportunity to determine the speed and direction of the target, and thus to use movement judgment, which was the construct we intended to measure. Therefore, the test was modified so that all targets are visible on the screen at the beginning of the trial and so that the subject is given at least a couple seconds to view the speed and direction of the target before the target reaches the optimal fire point.

Two modifications were made to Number Memory to reduce test administration time. The item part delay period was made a constant (1 second) rather than treated as a parameter with two levels (0.5 and 2.5 seconds), and the item string length (number of parts in an item) was changed from 4, 6, or 8 parts to 2, 3, or 4 parts. These changes drastically reduced the time required to complete the test. As a result, no reduction in the number of items, as had been recommended (see Table 9.2) was necessary. The Trial Battery version of this test had 28 items, constructed so that there were 13 replications of the four arithmetic operations (add, subtract, multiply, and divide).

Identical kinds of changes were made to the Target Tracking 1 and Target Tracking 2 tests. Internal consistency and test-retest reliability estimates were relatively high for these tests (internal consistency = .97 for both, test-retest = .68 and .77, for Tests 1 and 2, respectively), so we felt confident we could reduce the number of items from 27 to 18 in order to save time.

The difficulty of the test items was increased by increasing the speed of the crosshair and the target. This was done because field test data indicated that the Mean Distance Score was positively skewed; thus, the items appeared not to be differentiating very well among high ability subjects. By increasing the difficulty of the items, we hoped to create a more normal distribution of scores. Related to this, we used the ratio of target to crosshair speed as a test parameter, rather than target speed. It seemed to make sense that, given a particular crosshair speed, the ratio would be a better indicator of item difficulty than the actual target speed.

Finally, we modified the software controlling test administration so that the crosshair could not travel off the screen. During the field test, if a subject moved his/her crosshair so that it traveled off the screen (a not infrequent occurrence when the target was near the edge of the screen), he/she would lose sight of the crosshair. This caused problems for some subjects, who seemed not to know what to do when this happened.

Several changes were made to the Target Shoot Test. First, all test items were classified according to three parameters: crosshair speed, ratio of target to crosshair speed, and item complexity (i.e., number of turns/mean segment length). Then, items were revised in order to achieve a balanced number of items in each cell when the levels of these parameters were crossed. This had the result of "un-confounding" these parameters so that analyses could be made to see which parameters contributed to item difficulty.

Second, extremely difficult items were eliminated and item presentation times (the time the target was visible on the screen) were increased to a minimum of about 6 seconds (and a maximum of 10 seconds). This was done to eliminate a severe missing data problem for such items (as much as 40%), discovered during field tests. Missing data occurred when subjects failed to "fire" at a target. Time-to-Fire and Distance From Target scores could not be computed in these cases. These "no-fires" were found to occur where the target moved very rapidly or made many sudden changes in direction and speed, or the item lasted only a few seconds. Thus, the elimination of such items and increase in item time were intended to obviate the missing data problem. To save testing time, the number of items was reduced from 40 to 30, primarily by eliminating the extremely easy items. (Although test-retest reliabilities were only .48 and .58 for Mean Time-to-Fire and Mean Log Distance scores, respectively, we thought that solving the missing data problem would allow us to reduce the absolute number of items and still maintain this level of test-retest reliability.)

Finally, we added a feedback message to this test that reminded the subjects to press the red button (or "fire") when they had the crosshairs on the target, _if_ the subject failed to do so on the first practice item. This was done because a small percentage of subjects in the field test did not read the instructions carefully and treated this as a tracking test, i.e., they did not "fire" at the target until several items had been attempted. Usually the test administrator noticed this lapse by subjects, but placing this feedback message gave greater assurance that subjects would complete the test properly.

Changes to Non-Cognitive Measures (ABLE and AVOICE)

Table 9.4 presents a summary of the item-reduction changes that were made from the Pilot Trial Battery to the Trial Battery versions of the ABLE and AVOICE, as projected in Table 9.3. We needed to effect a 26 percent decrease in the total number of ABLE and AVOICE items. The goal in this revision was to decrease items on a scale-by-scale basis, while preserving the basic heterogeneity of each scale. The strategy adopted to accomplish this was as follows for each scale:

1. Sort items into content categories.

2. Rank order within category, based on item-scale correlations.

3. Drop last item in each category until desired number of items for that scale had been deleted.

Table 9.5 lists the ABLE scales and the number of items in each for the Pilot Trial Battery version and for the subsequent Trial Battery. Overall, the ABLE was decreased from 270 items to 209 items. In addition to deleting items, we standardized all response options on the ABLE by (1) changing the several four- and five-option responses to three-option responses and (2) ordering the response options so that the "highest" or "most" option (e.g., "All of the time") appeared first, and the "lowest" or "least" option (e.g., "None of the time") appeared third. Also, one last check was made to see whether there were still any response options that had such low endorsement rates as to be useless. A few such items were found, and their response options slightly modified.

AVOICE scale revisions are listed in Table 9.6. The total number of AVOICE items was decreased from 309 to 214. Thirty-eight of these 214 are items on the Work Environment Preference scales. It was decided to take this whole section out of the AVOICE booklet and include it in one of the criterion measure booklets, where a bit more administration time was available. Thus, 176 items remained in the AVOICE booklet.

As can be seen in Table 9.6, the decision was made to delete the Agriculture scale, the six single-item Holland scales, and the eight Expressed Interest items. There were no particularly compelling technical or psychometric reason for eliminating these scales; again, it was primarily a pragmatic decision in order to reduce the time necessary to complete the inventory. Reductions made on the remaining AVOICE scales were accomplished using the same strategy as that for the ABLE, decreasing scale length while preserving heterogeneity. The only items that had fewer than five response options were deleted in the above-described revisions, so the resultant Trial Battery AVOICE was made up entirely of five-option responses, from "Like Very Much" to "Dislike Very Much."

Table 9.4

Summary of Item Reduction Changes for ABLE and AVOICE

	No. of Items in PTB	No. of Items Recommended for Trial Battery[a]	No. of Items in Trial Battery
ABLE	270	199	209
AVOICE, excluding Organizational Climate/Environment Scales	269	188	176
AVOICE, Organizational Climate/Environment Scales	40	40	38
Total	579	427	423

---

[a] Based on IPR and SAG meetings described earlier in this chapter and summarized in Table 9.3.

Table 9.5

Number of Items in Pilot Trial Battery and Trial Battery Versions of ABLE Scales

ABLE Scale	No. of Items in PTB	No. of Items in Trial Battery
Emotional Stability	29	18
Self-Esteem	15	12
Cooperativeness	24	18
Conscientiousness	21	15
Nondelinquency	24	20
Traditional Values	16	11
Work Orientation	27	19
Internal Control	21	16
Energy Level	25	21
Dominance	16	12
Physical Condition	9	6
Adventure	8	8
Unlikely Virtues (Social Desirability)	12	12
Self-Knowledge	13	13
Non-Random Responses	8	8
Poor Impression	24	23
ABLE Total[a]	270	209

---

[a] This figure is not the simple sum of the number of items in each scale, since some items (e.g., on the Poor Impression Scale) are scored on more than one scale.

Table 9.6

Number of Items in Pilot Trial Battery and Trial Battery Versions of AVOICE Scales

AVOICE Scale	No. of Items in PTB	No. of Items in Trial Battery
Marksman	5	5
Agriculture	5	0
Mathematics	5	5
Aesthetics	5	5
Leadership	6	6
Electronic Communication	7	6
Automated Data Processing	7	6
Teacher/Counseling	7	6
Drafting	7	6
Audiographics	7	5
Armor/Cannon	8	7
Vehicle/Equipment Operator	10	6
Outdoors	9	9
Infantry	10	9
Science/Chemical Operations	11	7
Supply Administration	13	7
Office Administration	16	10
Law Enforcement	16	9
Mechanics	16	10
Electronics	20	12
Heavy Combat/Construction	23	13
Medical Services	24	12
Food Service	17	11
Adventure	6	6
Single-Item Holland Scales	6	0
Expressed Interest	8	0
Organizational Climate/ Environment Preferences	40	38 (moved to criterion booklet)
AVOICE Total[a]	309	214

---

[a] This figure is not the simple sum of the number of items in each scale since some items (e.g., on the Adventure Scale) are scored on more than one scale.

DESCRIPTION OF THE TRIAL BATTERY AND SUMMARY COMMENTS

In this chapter we have described the revisions made to the Pilot Trial Battery that produced the Trial Battery. In essence, the Trial Battery is a shortened and improved version of the Pilot Trial Battery used in the field tests. The Trial Battery was designed to be administered in a period of 4 hours and will be used during the Concurrent Validation phase of Project A.

Figure 9.2 shows a general description of the Trial Battery. These are the measures that were the product of the revisions just described. Appendix H contains copies of the Trial battery measures (Appendix H is in a separate limited-distribution report, ARI Research Note 87-24, as noted on p. xiv).

As already noted, the Trial Battery's intended use is as a predictor battery in the Concurrent Validation phase of Project A. Those data will allow the replication of analyses described here on a much larger sample (approximately 10,000). In addition, job performance criterion data will be collected which will allow an examination of the validity of Trial Battery measures for predicting soldiers' job performance. All of this information will be used to make revisions to the Trial Battery, thereby producing the Experimental Battery that will be used in a Longitudinal Validation effort in 1986 and later years. (See Figure 1.2 for a flow chart showing the relationships between the Pilot Trial Battery, Trial Battery, and Experimental Battery.)

Whatever the outcome of those future efforts, we think the development, pilot testing, and field testing leading up to the Trial Battery has reached the intended objectives. As already noted in Chapter 1 (see *Task 2: Progress Summary*), the measures developed came from a careful, structured process that identified the "best bets" for improving the prediction of soldiers' job performance. The new measures were developed using an iterative process that resulted in steady improvements guided by data. Procedures for efficiently and effectively administering the measures were developed along with the measures themselves. Finally, careful scrutiny of the psychometric characteristics of the measures shows them to be satisfactory to excellent in that regard.

COGNITIVE PAPER-AND-PENCIL TESTS

Name	Number of Items	Time Limit
Reasoning Test	30	12 minutes
Object Rotation Test	90	7.5 minutes
Orientation Test	24	10 minutes
Maze Test	24	5.5 minutes
Map Test	20	12 minutes
Assembling Objects Test	32	16 minutes

PERCEPTUAL/PSYCHOMOTOR COMPUTER-ADMINISTERED TESTS

Name	Number of Items	Approximate Time
Demographics	2	4 minutes
Reaction Time 1	15	2 minutes
Reaction Time 2	30	3 minutes
Memory Test	36	7 minutes
Target Tracking Test 1	18	8 minutes
Perceptual Speed and Accuracy Test	36	6 minutes
Target Tracking Test 2	18	7 minutes
Number Memory Test	28	10 minutes
Cannon Shoot Test	36	7 minutes
Target Identification Test	36	4 minutes
Target Shoot Test	30	5 minutes

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES

Name	Number of Items	Approximate Time
Assessment of Background and Life Experiences (ABLE)	209	35 minutes
Army Vocational Interest Career Examination (AVOICE)	176	20 minutes

Figure 9.2. Description of Trial Battery measures.

APPENDIX A

Data Bases Searched

PSCYINFO. (Commonly known as Psyc Abstracts) This file is produced by the American Psychological Association and covers the world's literature in psychology and related behavioral and social sciences such as psychiatry, sociology, anthropology, education, pharmacology, and linguistics. The following general fields are covered: applied psychology, educational psychology, experimental human and animal psychology, experimental social psychology, general psychology, personality, physical and psychological disorders, physiological intervention, physiological pathology, professional personnel and issues, psychometrics, social processes and issues, treatment and prevention.

GPOM. (Government Printing Office Monthly Catalog) This file is produced by the Superintendent of Documents, United States Government Printing Office and indexes the public documents generated by the legislative branch, executive branch, and all agencies of the United States Federal Government. Some publications from the judicial branch are also included. The subjects covered are agriculture, commerce, defense, health and human services, education energy, housing, interior, justice, labor, state, transportation, and treasury.

NTIS. (National Technical Information Service) This file is produced by the National Technical Information Service of the U.S. Department of Commerce. The data base consists of government-sponsored research, development, and engineering reports as well as other analyses prepared by government agencies, their contractors, or grantees. The following are representative of the subject areas: administration and management; aeronautics and aerodynamics; agriculture and food; astronomy and astrophysics; atmospheric sciences; behavior and society; biomedical technology and engineering; building industry technology; business and economics; chemistry; civil engineering; communication; computers, control, and information theory; electrotechnology; energy; environmental pollution and control; health planning; industrial and mechanical engineering; library and information sciences; materials sciences; mathematical sciences; medicine and biology; military sciences; missile technology; natural resources and earth sciences; navigation, guidance, and control; nuclear science and technology; ocean technology and engineering; photography and recording devices; physics; propulsion and fuels; space technology; transportation; urban and regional technology.

ERIC. (Educational Resources Information Center) This data file is produced by The National Institute of Education and covers the following subject areas: adult, career, and vocational education; counseling and personnel services; early childhood education; educational management; handicapped and gifted children; higher education; information resources; junior colleges; languages and linguistics; reading and communication skills; rural education and small schools; science, mathematics, and environmental education; social studies/social science education; teacher education; tests, measurement, and evaluation; and urban education.

SSCI & SSCB. (Social Scisearch) These files are produced by the Institute for Scienfific Information (ISI) and constitute an international, multidisciplinary index to the literature of the social, behavioral, and related sciences. Subjects included in the data base are anthropology, archaeology, area studies, business and finance, communication, community health, criminology and penology, demography, economics, education research, ethnic group studies, geography, history, information/library science, international relations, law, linquistics, management, marketing, philosophy, political science, psychology, psychiatry, sociology, statistics, and urban planning and development.

SSIE. (Smithsonian Science Information Exchange) This file is produced by the Smithsonian Science Information Exchange and contains abstracts of research either in progress or completed in the past two years. The data bases encompass all fields of basic and applied research in the physical, social, engineering, and life sciences including: agricultural sciences, behavioral sciences, biological sciences, chemistry and chemical engineering, earth sciences, electronics, engineering materials, mathematics, medical sciences, physics, social sciences and economics.

DTIC. (Defense Technical Information Center) This file is produced by the Defense Logistics Agency. It makes available from one central repository the thousands of research and development reports produced each year by U.S. military organizations and their contractors and grantees. Defense facilities and their contractors are required to submit to DTIC copies of each report (up to and including SECRET) that formally records scientific and technical results of Defense-sponsored research, development, test, and evaluation. Although created originally to serve the military, DTIC services habe been extended to all federal government agencies and their contractors, subcontractors, and grantees.

APPENDIX B

Copies of Article Review and Predictor Review Forms

A-__-___
Article Code

ARTICLE REVIEW FORM

Reviewer Initial

A. CITATION

☐ Article ☐ Book/Monograph ☐ Text Manual ☐ Technical Report ☐ Others

☐ Check here if not reviewed; explain why below:

B. ABSTRACT

C. LIST OF PREDICTORS

Predictor Rev Form Codes:

B-2

D. DESCRIPTION OF CRITERION MEASURES

CRITERION 1

☐ Job Proficiency ☐ Training Performance ☐ Other

Description:

Development:

Reliability: Value(s) ____________________ Type and Method of Estimation:	Descriptive Statistics (N, $\bar{x}$, S.D.):

CRITERION 2

☐ Job Proficiency ☐ Training Performance ☐ Other

Description:

Development:

Reliability: Value(s) ____________________ Type and Method of Estimation:	Descriptive Statistics (N, $\bar{x}$, S.D.):

E - SAMPLES

SAMPLE 1

Purpose: ______________________________

Description:

Race

		W	B	Asian	Hisp	Am Ind	Other	Total
Sex:	**M**							
	F							
	Total							

Age: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

Educ: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

(Explain, if scale:)

SAMPLE 2

Purpose: ______________________________

Description:

Race

		W	B	Asian	Hisp	Am Ind	Other	Total
Sex:	**M**							
	F							
	Total							

Age: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

Educ: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

(Explain, if scale:)

SAMPLE 3

Purpose: ______________________________

Description:

Race

		W	B	Asian	Hisp	Am Ind	Other	Total
Sex:	**M**							
	F							
	Total							

Age: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

Educ: $\bar{x}$ ______________ **S.D.** ______________ **Range** ______________ **Median** ______________

(Explain, if scale:)

F. METHODOLOGY

Methodology: Check all that apply.

- ☐ Criterion related: concurrent
- ☐ Criterion related: predictive
- ☐ Content validity
- ☐ Factor Analytic or Psychometric
- ☐ Reanalysis, review, or summary of data or past studies
- ☐ Other: ____________________

Details of Methodology:

G. OTHER RESULTS

H. REVIEWER'S COMMENTS

Opinions about research design, etc.

_-__-____

Predictor Code

Reviewer Initials

PREDICTOR REVIEW FORM

Predictor Title: ____

Construct (Taxon): ____

Intended to measure: ____

Brief description of predictor: ____

Description of items/tasks:

A. DESCRIPTION OF PREDICTOR

Number of items/trials: ____

Power/Speeded/ ____ Time Limit/Approx. Time ____

Administration Procedures:

Scoring Procedures:

Publisher Clue:

Article Code: A - ____

B. PREDICTOR RELIABILITY

Value	N	Type	Method of Estimation

C. NORMS/DESC STATS

Mean	S.D.	N	Group Description

D. CORRELATIONS WITH OTHER PREDICTORS

r	N	Predictor Description

(Describe sample(s), circumstances, etc., if necessary.)

E. CORRELATIONS WITH CRITERIA

r	N	Criterion Measure	Sample

(Describe sample, etc., if necessary.)

F. ADVERSE IMPACT/DIFFERENTIAL VALIDITY/TEST FAIRNESS

G. RECOMMENDATIONS

APPENDIX C

Names and Definitions of Predictor and Criterion Variables Used in Expert Judgment Task

List of 53 Predictor Variables Identified For Inclusion in the Expert Judgment Task

PREDICTOR VARIABLES

Construct Name	Definition
Verbal Comprehension	Measures knowledge of the meaning of words and their relationships to each other.
Numerical Computation	Measures speed and accuracy in performing simple arithmetic operations, i.e., addition, subtraction, multiplication, and division.
Use of Formulations and Number Problems	Measures the ability to correctly use algebraic formulae to solve number problems.
Word Problems	Measures the ability to select and organize relevant information to correctly solve mathematical word problems.
Reading Comprehension	Measures the ability to read and understand written material.
Two-Dimensional Mental Rotation	Measures the ability to identify a two-dimensional figure when seen at different angular orientations within the picture plane.
Three-Dimensional Mental Rotation	Measures the ability to identify a three-dimensional object, projected on a two-dimensional plane, when seen at different angular orientations either within the picture plane or about the axis in depth.
Inductive Reasoning: Concept Formation	Measures the ability to discover a rule or principle and apply it in solving a problem.
Spatial Visualization	Measures the ability to mentally manipulate the components of a two- or three-dimensional figure into other arrangements.
Deductive Logic	Ability to use logic and judgment in drawing conclusions from available information. Given a test of facts and a set of conclusions, deductive logic refers to the ability to determine whether the conclusions flow logically from the facts.
Field Dependence	Ability to find a simple form when it is hidden in a complex pattern. Given a visual percept or configuration, field dependence (or independence, more accurately) refers to the ability to hold it in mind so as to disembed it from other well-defined perceptual material.
Perceptual Speed and Accuracy	Ability to perceive visual information quickly and accurately and to perform simple processing tasks with it (e.g., comparisons). This requires the ability to make rapid scanning movements without being distracted by irrelevant visual stimuli, and also measures memory, working speed, and sometimes eye-hand coordination.

PREDICTOR VARIABLES

Construct Name	Definition
Mechanical Comprehension	Ability to learn, comprehend, and reason with mechanical terms. More specifically, this is the ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations.
Rote Memory	Measures the ability to recall previously learned but unrelated item pairs.
Place Memory (Visual Memory)	Ability to remember the configuration, location, and orientation of figural material.
Ideational Fluency	Ability to rapidly generate ideas about a given topic or exemplars of a class of objects.
Follow Directions	Measures ability to follow simple and complex directions.
Analogical Reasoning	Measures the ability to identify the underlying principles governing relationships between pairs of objects.
Figural Reasoning	Measures ability to generate and apply hypotheses about principles governing the relationship among several figures.
Spatial Scanning	Measures the ability to visually survey a complex field to find a particular configuration representing a pathway through the field.
Omnibus Measures of Intelligence/Aptitude	Measures general mental ability or general attitude.
Word Fluency	Ability to rapidly think of words.
Verbal and Figural Closure	Measures ability to identify objects or words given sketchy or partial information.
Processing Efficiency	Speed of reactions to simply stimuli.
Selective Attention	This is the ability to attend to a target stimulus when presented with two or more stimuli simultaneously.
Time-Sharing	Time-sharing is the ability to perform two or more tasks simultaneously.
Multilimb Coordination	Multilimb coordination is the ability to coordinate the simultaneous movement of two or more limbs. This ability is general to tasks requiring coordination of any two limbs (e.g., two hands, two feet, one foot and one hand). It is most common to tasks where the body is at rest (e.g., seated or standing) while two or more limbs are in motion.

PREDICTOR VARIABLES

Construct Name	Definition
Control Precision	Control precision is the ability to make fine, highly controlled (but not over-controlled) muscular movements necessary to adjust or position a machine or equipment control mechanism. This ability is general to tasks requiring motor adjustments in response to a stimulus whose speed and/or direction of movement are perfectly predictable. This ability is critical in situations where the motor adjustments must be both rapid and precise. The ability extends to arm-hand movements as well as to leg movements.
Rate Control	Rate control is the ability to make continuous anticipatory muscular movements necessary to adjust or position a machine or equipment control mechanism. This ability is general to tasks requiring motor adjustments or movements in response to a moving stimulus which is changing speed and/or direction in a random or unpredictable manner. The ability applies to compensatory tracking of the stimulus as well as following pursuit of the stimulus.
Manual Dexterity	Manual dexterity is the ability to make skillful, coordinated movements of the hand or the arm and hand. This ability most typically applies to tasks involving manipulation of moderately large objects (e.g., blocks, pencils, etc.) under speeded conditions.
Finger Dexterity	Finger dexterity is the ability to make skillful, coordinated, highly controlled movements of the fingers. This ability applies primarily to tasks involving manipulation of objects with the fingers.
Track Tracing Test	Designed to measure arm-hand steadiness.
Wrist-Finger Speed	The ability to carry out very rapid, discrete movements of the fingers, hands, and wrists. This ability applies primarily to tasks in which the accuracy of the movement is not a major concern. This ability is determined entirely by the speed with which the movement is carried out.
Aiming	The ability to make very precise, accurate hand movements under highly-speeded conditions. This ability is dependent upon very precise eye-hand coordination.
Speed of Arm Movement	This ability involves the speed with which discrete arm movements can be made. The ability deals with the speed with which the movement can be carried out after it has been initiated.

PREDICTOR VARIABLES

Construct Name	Definition
Involvement in Athletics and Physical Conditioning	Frequency and degree of participation in sports, exercise, and physical activity. Individuals high on this dimension actively participate in individual and team sports and/or exercise vigorously several times per week.
Energy Level	Characteristic amount of energy and enthusiasm. The person high in energy level is enthusiastic, active, vital, optimistic, cheerful, zesty, and has the energy to get things done.
Cooperativeness	Characteristic degree of pleasantness versus unpleasantness exhibited in interpersonal relations. The highly cooperative person is pleasant, tolerant, tactful, helpful, not defensive, and generally easy to get along with. His/her participation in a group adds cohesiveness.
Sociability	Outgoingness. The person high in sociability is talkative, relates easily to others, is responsive and expressive in social environments, readily becomes involved in group activities, and has many relationships.
Traditional Values	Personal views in areas such as authority, discipline, social change, and religious commitment. The person with traditional values accepts authority and the value of discipline, is likely to be religious, values propriety, and is conventional, conservative, and resistant to social change.
Dominance	Tendency to seek and enjoy positions of leadership and influence over others. The highly dominant person is forceful and persuasive at those times when adopting such characteristics is appropriate.
Self-esteem	Degree of confidence in one's abilities. A person with high self-esteem feels largely successful in past undertakings and expects to succeed in future undertakings.
Conscientiousness	Characteristic amount of behavioral self-control. The highly conscientious person is dependable, planful, well organized, and disciplined. This person prefers order and thinks before acting.
Locus of Control	Characteristic belief in the amount of control people have over rewards and punishments. The person with an internal locus of control expects that there are consequences associated with behavior and that people control what happens to them by what they do. The person with an external locus of control believes that what happens to people is beyond their personal control.

PREDICTOR VARIABLES

Construct Name	Definition
Emotional Stability	Characteristic degree of stability vs. reactivity of emotions. The emotionally stable person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He/she thinks clearly and maintains composure and rationality in situations of actual or perceived stress.
Nondelinquency	Amount of respect for laws and regulations as manifested in attitudes and behavior. The nondelinquent person is honest, trustworthy, wholesome, and law-abiding. Such persons will have histories devoid of trouble with schools and legal agencies.
Work Orientation	Tendency to strive for competence in one's work. The work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in completion of the task at hand.
Realistic Interests	Preference for concrete and tangible activities, characteristics, and tasks. Persons with realistic interests enjoy, and are skilled in, the manipulation of tools, machines, and animals, but find social and educational activities and situations aversive.
Investigative Interests	Preference for scholarly, intellectual, and scientific activities and tasks. Persons with investigative interests enjoy analytical, ambiguous, and independent tasks, but dislike leadership and persuasive activities.
Enterprising Interests	Preference fo persuasive, assertive, and leadership activities and tasks. Persons with enterprising interests may be characterized as ambitious, dominant, sociable, and self-confident.
Artistic Interests	Preferences for unstructured, expressive, and ambiguous activities and tasks. Persons with artistic interests may be characterized as intuitive, impulsive, creative, and non-conforming.
Social Interests	Preferences for social, helping, and teaching activities and tasks. Persons with social interests may be characterized as responsible, idealistic, and humanistic.
Conventional Interests	Preferences for well-ordered, systematic, and practical activities and tasks. Persons with conventional interests may be characterized as conforming, unimaginative, efficient, and calm.

CRITERION CONSTRUCTS

1. Inspect mechanical systems--test, measure, and/or use diagnostic equipment as well as visual, aural and tactile senses, in conjunction with technical information, to compare the operating status of mechanical equipment (e.g., engines, transmissions, machineguns) and mechanical components (e.g., bearings in an electrical generator) to standards of operating efficiency, and to identify malfunctions.

   Actions may include: analyze, read, operate

2. Troubleshoot mechanical systems--use test, measuring, and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in mechanical equipment (e.g., engines, transmissions, machineguns) and mechanical components (e.g., bearings in an electrical generator).

   Actions may include: analyze, read, calculate

3. Repair mechanical systems--perform corrective actions on previously diagnosed malfunctions of mechanical equipment or mechanical components using appropriate tools (e.g., wrenches, screwdrivers, gauges, hammers) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read, work metal

4. Inspect fluid systems--use test, measuring, and diagnostic equipment, as well as visual, aural and tactile senses, in conjunction with technical information, to determine the operating status of fluid systems (e.g., hydraulic, refrigeration, engine cooling, compressed air) in comparison to standards of operating efficiency, and to identify malfunctions.

   Actions may include: analyze, read, operate

5. Troubleshoot fluid systems--use test, measuring and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in fluid systems (e.g., hydraulic, refrigeration, engine cooling, compressed air).

   Actions may include: analyze, read, calculate

6. Repair fluid systems--perform corrective actions on previously diagnosed malfunctions of fluid systems using appropriate tools (e.g., wrenches, pressure gauges, soldering equipment) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read

7. Inspect electrical systems--use test, measuring, and diagnostic equipment, as well as visual, aural and tactile senses, in conjunction with technical information, to determine the operating status of electrical systems (e.g., generators, wiring harnesses, switches, relays, circuit breakers, motors, lights) in comparison to standards of operating efficiency and to identify malfunctions.

   Actions may include: Analyze, read, operate

8. Troubleshoot electrical systems--use test, measuring and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in electrical systems (e.g., generators, wiring harnesses, switches, relays, circuit breakers, motors, lights).

   Actions may include: analyze, read, calculate

9. Repair electrical systems--perform corrective actions on previously diagnosed malfunctions of electrical systems and electrical components using appropriate tools (e.g., pliers, wire strippers, soldering irons) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read

10. Inspect electronic systems--use test, measuring and diagnostic equipment, and to a limited extent, visual, aural, and tactile senses, in conjunction with technical information, to compare the operating status of electronic systems (e.g., communications equipment, radar, missile and tank ballistics controls) to standards of operating efficiency and to identify malfunctions.

    Actions may include: analyze, read, operate

11. Troubleshoot electronic systems--use test, measuring, and diagnostic equipment, in conjunction with technical information, to determine the cause or location of malfunctions in electronics systems (e.g., communication equipment, radar, missile and tank ballistics controls).

    Actions may include: analyze, read, calculate

12. Repair electronic systems--perform corrective actions on previously diagnosed malfunction of electronic systems and electronic components using appropriate tools (e.g., test sets, screwdrivers, pliers, soldering guns) in conjunction with technical information.

    Actions may include: adjust, assemble/disassemble, install, fix, read

13. Repair metal—perform corrective actions (e.g., bend, cut, drill, saw, weld, rivet, hammer, grind, solder, paint) to refabricate metal structures.

    Actions may include: calculate, assemble/disassemble, fix, construct, read, work metal

14. Repair plastic and fiberglass structures--perform corrective actions (e.g., measure, cut, saw, drill, sand, fill, paint, glue) to refabricate plastic and fiberglass structures.

    Actions may include: calculate, assemble/disassemble, fix, construct, read

15. Construct wooden buildings and other structures--perform carpentry activities (e.g., measure, saw, nail, plane) to frame, sheath and roof buildings, or to erect trestles, bridges, piers, etc.

    Actions may include: calculate, assemble/disassemble, install, construct, read

16. Construct masonry buildings and structures--perform masonry activities (e.g., measure, lay brick, pour concrete) to construct walls, columns, field fortifications, etc.

    Actions may include: construct, calculate, assemble/disassemble, read

17. Prepare parachutes--inspect cargo and personnel parachutes, repair or replace faulty parachute components, and prepare (i.e., pack) parachute for future air drop.

    Actions may include: adjust, assemble/disassemble, pack/unpack, fix, sew, read

18. Prepare equipment and supplies for air drop--fabricate and assemble platforms, cushions, and rigging to parachute supplies, equipment and vehicles; load, position and secure supplies and equipment in aircraft.

    Actions may include: adjust, assemble/disassemble, pack/unpack, construct, transport

19. Install electronic components--place and interconnect electronic and communication components and equipment (e.g., radios, antennas, telephones, teletypewriters, radar, power supplies) and check system for operation.

    Actions may include: adjust, assemble/disassemble, install, read

20. Operate electronic equipment--set and adjust the controls of electronic components to operate electronic systems (e.g., radio, radar, computer hardware, missile ballistics controls).

Actions may include: adjust, operate

21. Send and receive radio messages--use standardized radio codes and procedures to transmit and receive information.

Actions may include: signal, communicate, read

22. Operate keyboard device--type information using a typewriter, teletype or keypunch, or computer terminal.

Actions may include: process, operate

23. Use maps in the field--read and interpret map symbols and identify geography features in order to locate geography features and field positions on the map, and to locate map features in the field.

Actions may include: analyze, identify, read, calculate

24. Plan placement or use of tactical position and features--using maps and on-site inspection, identify geographic positions or areas to be used for cover and concealment or to place fortifications, mines, detectors, chemicals, etc.

Actions may include: analyze, calculate, read

25. Place tactical equipment and materials in the field--without using heavy equipment (e.g., lifts, dozers), place mines, detectors, chemicals, camouflage or other tactical items into position on the battlefield.

Actions may include: use weapons, maneuver, transport, install

26. Detect and identify targets--using primarily sight, with or without optical systems, locate potential targets, and identify type (e.g., tanks, troops, artillery) and threat (friend or foe); report information.

Actions may include: communicate, analyze

27. Prepare heavy weapons for tactical use--transport, position and assemble heavy tactical weapons such as missiles, field artillery, anti-aircraft systems.

Actions may include: adjust, assemble/disassemble, install, pack/unpack

28. Load field artillery or tank guns--manipulate breech controls and handle ammunition (stow and load) to prepare guns for firing.

Actions may include: use weapons, pack/unpack

29. Fire heavy direct fire weapons (e.g., tank main guns, TOW missile, infantry fighting vehicle cannon)--using optical sighting systems, manipulate weapon system controls to aim, track and fire on designated targets.

Actions may include: use weapons, operate, adjust

30. Operate fire controls of indirect fire weapons (e.g., field artillery)--using map coordinates and ballistics information determine elevation and azimuth needed for firing at designated targets; adjust weapon using fire controls.

Actions may include: analyze, calculate, read, adjust

31. Fire individual weapons--aim, track and fire hand operated weapons such as rifles, pistols, and machineguns at designated targets.

Actions may include: use weapons

32. Engage in bayonet and hand-to-hand combat--use offensive and defensive body maneuvers to subdue hostile individuals.

Actions may include: maneuver, apprehend

33. Operate wheeled vehicles--use various vehicle controls to drive wheeled vehicles from point to point, generally over paved and unpaved roads, observe traffic regulations; secure cargo.

Actions may include: maneuver, transport, operate

34. Operate track vehicles--use various vehicle controls to drive track vehicles (e.g., tanks, APCs, scout vehicles, bulldozers); steer in response to terrain features.

Actions may include: maneuver, transport, operate

35. Operate lifting, loading and grading equipment--operate heavy equipment (e.g., fork lifts, cranes, loader, back-hoes, graders) to load, unload, or move heavy equipment, supplies, construction materials (e.g., culvert pipes, building or bridge trusses), or terrain features (e.g., earth, rock, trees).

Actions may include: construct, operate

36. Operate power excavating equipment--use pneumatic hammers and drills, paving breakers, grinders, and backfill tampers, in the fabrication and modification of concrete, stone and earthen structures.

    Actions may include: construct, operate

37. Reproduce printed materials--operate duplicating machines and offset presses to reproduce printed materials; collate and bind materials using various types of bindery equipment.

    Actions may include: adjust, operate, photograph, calculate

38. Make movies and videotapes--use motion picture cameras or videotape equipment to record visual and auditory aspects of assigned subject matter to be used for intelligence analyses, training or documentation.

    Actions may include: adjust, photograph

39. Draw maps and overlays--use drafting, graphics, and related techniques to prepare and revise maps, with symbols and legends, from aerial photographs.

    Actions may include: analyze, process, draw

40. Write and deliver presentations--prepare scripts for formal presentation including radio and television broadcast; make oral presentations.

    Actions may include: analyze, write

41. Record and file information--collect, transcribe, annotate, sort, index, file, and retrieve information (e.g., training rosters, personnel statistics, supply inventories).

    Actions may include: process, dispose

42. Receive, store and issue supplies, equipment and other materials--inspect material and review paperwork upon receipt; sort, transport, and store material; issue or ship material to authorized personnel or units.

    Actions may include: analyze, calculate, process, send, pack/unpack, transport

43. Prepare technical forms and documents--follow standardized procedures to prepare or complete forms and documents (e.g., personnel records and dispositions, efficiency reports, legal briefs).

    Actions may include: process, write, analyze

44. Translate or decode data--use standardized coding systems and decoding rules to convert coded information to some more usable form (e.g., interpret radar information, decode Morse code, translate foreign languages).

    Actions may include: analyze

45. Analyze intelligence data--determine importance and reliability of information; integrate information to provide identification, disposition and movement of enemy forces and estimate enemy capabilities.

    Actions may include: communicate, analyze, read

46. Prepare food--prepare food and beverages according to recipes and meal plans (measure, mix, bake, etc.); inspect fresh food and staples for freshness; maintain sanitary work area.

    Actions may include: cook, read, sanitize, dispose, calculate

47. Receive clients, patients, guests--schedule, greet and give routine information to persons seeking medical, dental, legal or counseling services.

    Actions may include: administer, communicate, process

48. Interview--verbally gather information from clients, patients, witnesses, prisoners, or other persons.

    Actions may include: communicate

49. Provide medical and dental treatment--give medical attention to soldiers in the field, or medical or dental clinic, or to animals (e.g., CPR, splinting fractures, administering injections, dressing wounds).

    Actions may include: treat, sanitize, photograph

50. Select, lay-out and clean medical or dental equipment and supplies--prepare treatment areas for use by following prescribed procedures for laying-out instruments and equipment; clean equipment and area for subsequent use.

    Actions may include: sanitize, assemble/disassemble, pack/unpack, dispose

51. Perform medical laboratory procedures--conduct various types of blood tests, urinalysis, cultures, etc.

    Actions may include: sanitize, analyze, calculate, adjust

52. **Control individuals and crowds--apprehend suspected criminals, capture enemy soldiers, guard prisoners, participate in riot control operations, etc.**

Actions may include: apprehend, communicate, administer

53. **Control air traffic--coordinate departing, en route, arriving and holding aircraft by monitoring radar equipment and communicating with aircraft and other air traffic control facilities.**

Actions may include: communicate, analyze, send, operate, signal

Initial Training Performance Variables

1. Training progress/success—successfully completing formal training course in normal amount of time versus washing out, being reassigned, being "set back" or "recycled."

2. Effort/motivation in training—the degree of effort, motivation, and interest that a soldier puts into his/her training, as evidenced by such things as curiosity about course content, not being afraid to be "wrong" or to ask questions, taking notes, being attentive in class, studying on own time, seeking out the instructor to clarify course content.

3. Performance of theoretical, or "classroom" parts of training—learning the theoretical part of a course; performing well on quizzes, tests, and examinations given in a classroom setting that tests the acquisition of concepts, principles, facts, or other information, e.g., learning the basic food groups, understanding the principles of internal combustion, learning the nomenclature of a weapon.

4. Performance of practical, "hands-on" part of training—applying the theory or principles of a course to practical problems and situations, either during simulations, field exercises, or other "hands-on" parts of training, e.g., cooking a meal, repairing an engine, firing a weapon, etc.

Nine Behavioral Dimensions of Generalized Army Effectiveness

1. Following regulations—consistently complying with Army rules and regulations; conforming appropriately to standard procedures; following the spirit as well as the letter of military and civilian laws, regulations, written orders, etc.

2. Commitment to Army norms—adjusting successfully to Army life; displaying appropriate military appearance and bearing; showing pride in being a soldier.

3. Cooperation with supervisors—responding willingly to orders, suggestions, and other guidance from NCOs and officers; deferring appropriately to superiors' expertise and judgment and being supportive of superior officers/NCOs.

4. Cooperation with other unit members—pitching in when necessary to help other unit members with their job and mission assignments or during training; encouraging and supporting other unit members, as appropriate; showing concern for unit objectives over and above personal interests.

5. Hard work and perseverance—working hard on the job and during training; sustaining maximum effort over long periods of hard duty and on daily assignments; coping well with hardship or otherwise unpleasant conditions to continue to work toward mission completion.

6. Attention to detail—carrying out assignments carefully and thoroughly; consistently completing job and duty assignments on time or ahead of schedule; being conscientious in maintaining own and unit's equipment, and taking care to ensure that own quarters are clean and neat.

7. Initiative—willingly volunteering for assignments; performing extra necessary tasks without explicit orders; anticipating problems and taking action to prevent them.

8. Discipline—consistently concentrating on the job or duty assignment rather than being distracted by opportunities to socialize or otherwise stop working; controlling own emotions and not allowing them to interfere with performance of duty; keeping under control alcohol and other drug intake so that performance is not affected.

9. Emergent leadership—displaying good judgment in making suggestions to others in the unit regarding the job, duty assignments, etc.; appropriately taking charge when placed in a leadership position; where appropriate, persuading others in the unit to accept his/her ideas, opinions, and directions.

Six General Army Effectiveness Variables

10. Survive in the field—react to direct or indirect fire; construct individual fighting position; camouflage self and equipment; use challenge and password; protect against NBC attack.

11. Maintain physical fitness—keep self at physical fitness level appropriate for state of battle readiness.

12. Disciplinary problems—having a record of disciplinary problems as reflected by AWOLS, Article 15s, civil arrests, etc.

13. Attrition—separating from the Army for "negative" reasons such as discipline or drug-related problems.

14. Reenlistment—signing on for a second tour of duty.

15. Job satisfaction/morale—being satisfied with own MOS and Army life.

APPENDIX D

Scale Names and Number of Items in Each Scale for the Preliminary Battery

Scale Names and Number of Items in Each Scale for the Preliminary Battery

The scale names, with the number of items each included parenthetically, are as follows:

Perceptual-cognitive: ETS Figure Classification (FC: 28 items with 8 responses each); ETS Map Planning (MP: 40); ETS Choosing a Path (CP: 32); ETS Following Directions (FD: 20); ETS Hidden Figures (HF: 32); EAS Space Visualization (SV: 50); EAS Numerical Reasoning (NR:20); Flanagan Assembly (FNA: 20).

Vocational interests (VOICE): Office Administration (20); Heavy Construction (20); Electronics (20); Medical Service (20); Outdoors (15); Aesthetics (15); Mechanics (15); Food Services (15); Law Enforcement (15); Agriculture (15); Mathematics (12); Audiographics (10); Teacher/Counseling (10); Marksman (7); Drafting (7); Craftman (7); Automated Data Processing (7).

Temperament (Personnel Opinion Inventory or POI): Conscientiousness (DPQ Unlikely Virtues/PRF Infrequency: 10); Leadership (DPQ Social Potency: 26); Stress (DPQ Stress Reaction: 26); Discipline (CPI Socialization: 30); Motivation (Rotter I/E Locus of Control: 29).

Biographical Questionnaire (BQ): Scales for Males. Warmth of Parental Relationship (11); Academic Achievement (25); Social Introversion (22); Athletic Interest (10); Intellectualism (18); Aggressive/Independence (10); Parental Control vs. Freedom (11); Social Desirability (10); Scientific Interest (12); Academic Attitude (8); Sibling Friction (5).

Scales for Females. Warmth of Maternal Relationship (13); Social Leadership (22); Academic Achievement (13); Parental Control vs. Freedom (11); Cultural Literary Interests (5); Athletic Participation (9); Scientific Interest (13); Feelings of Social Inadequacy (3); Adjustment (5); Expression of Negative Emotion (4); Social Maturity (2); Popularity with Opposite Sex (4); Positive Academic Attitude (7); Warmth of Parental Relationship (5).

Rational (Combined Sex) Scales: Leadership (12); Social Confidence (4); Social Activity (11); Self Control (5); Antecedents of Self Esteem (6); Parental Closeness (13); Sibling Harmony (5); Independence (8); Academic Confidence (5); Academic Achievement (6); Positive Academic Attitude (6); Effort (4); Scientific Interests (5); Reading/Intellectual Interests (6); Athletic Interests (2); Athletic/Sports Participation (6); Physical Condition (18); Vocational-Technical Activities (4).

APPENDIX E

Computerized Measures Observed During Site Visits for ARI Project A, Spring 1983

COMPUTERIZED MEASURES OBSERVED DURING SITE VISITS FOR ARI PROJECT A

	LOCATION					MACHINE					PROGRAMMING LANGUAGE		
PREDICTOR	AFHRL	NAMRL	FT. RUCKER	FT. KNOX		APPLE	TERAK	PDP 11	ANALOGUE APPARATUS		PASCAL	BASIC	FORTRAN
PERCEPTUAL													
Simple Reaction Time	✓						✓				✓		
Choice Reaction Time (2-6)	✓						✓				✓		
Posner Physical Identity	✓						✓				✓		
Posner Name Identity	✓						✓				✓		
Single Word Classification	✓						✓				✓		
Comparison of Word Prs.	✓						✓				✓		
Line Length Judgments	✓						✓				✓		
Visual Search	✓						✓				✓		
Rotated Figures	✓						✓				✓		
Perceptual Speed	✓							✓					✓
DOT Estimation	✓							✓					✓
Mental Rotation	✓							✓					✓
Decision Making Speed (CRT)	✓							✓					✓
Embedded Figures	✓							✓					✓
Card Rotation[1]		✓						✓					
Hidden Patterns[1]		✓						✓					
Maze Training[1]		✓						✓					
Perceptual Speed Test			✓			✓						✓	
INFORMATION PROCESSING													
Sternberg Numbers	✓						✓				✓		
Sternberg Words	✓						✓				✓		
Old-New Item Recognition	✓						✓				✓		
Random Two Responses	✓						✓				✓		
Nine Digit Short Term Memory	✓						✓				✓		
Continuous Paired Assoc.	✓						✓				✓		
Dual Task-Tapping & Visual	✓						✓				✓		
Visual Memory (5x5)	✓						✓				✓		
Time Sharing: Compensatory Tracking & Digit Cancellation	✓							✓					✓

---

[1] These measures administered under NAMRL contract at the Aviation Research Laboratory in Illinois.

COMPUTERIZED MEASURES OBSERVED DURING SITE VISITS FOR ARI PROJECT A

PREDICTOR	LOCATION				MACHINE				PROGRAMMING LANGUAGE		
	AFHRL	NAMRL	FT. RUCKER	FT. KNOX	APPLE	TERAK	PDP 11	ANALOGUE APPARATUS	PASCAL	BASIC	FORTRAN
INFORMATION PROCESSING (CONT.)											
Encoding Speed	✓						✓				✓
Immediate/Delayed Memory	✓						✓				✓
Item Recognition	✓						✓				✓
Time Sharing: Compensatory Tracking & Arithmetic[2]		✓						✓			
Selective Attention (DLT)[2]		✓	✓					✓			
Time Sharing: Stick & Rudder & DLT		✓						✓			
Sternberg Memory Search Tasks 1-4[3]		✓					✓				
Delayed Digit Recall[3]		✓	X		X		✓			X	
Time Sharing: Compensatory Tracking & CRT			✓		✓					✓	
COGNITIVE											
Numerical Operations	✓					✓			✓		
Sentence Verification	✓					✓			✓		
Paired Assoc. Learning	✓					✓			✓		
Moyer-Landauer Task	✓					✓			✓		
Relearning of Paired Assoc.	✓					✓			✓		
Three Term Comparisons	✓					✓			✓		
Similarity Judgments	✓					✓			✓		
Days of Week Addition	✓					✓			✓		
Simon-Kotovsky Task	✓					✓			✓		
Word-Nonword Comparison	✓					✓			✓		
Collins & Quillian	✓					✓			✓		
Adaptive Vocabulary	✓					✓			✓		
Thurstone's ABC	✓					✓			✓		
Risk Taking	✓						✓				✓
Word Knowledge	✓						✓				✓
M-1 Computer Panel Test				✓	✓					✓	

[2] NAMRL is in the process of adapting these to an Apple computer with joy stick, foot pedals and a speech generation chip.

[3] These measures administered under NAMRL contract at the Aviation Research Laboratory in Illinois.

COMPUTERIZED MEASURES OBSERVED DURING SITE VISITS FOR ARI PROJECT A

PREDICTOR	LOCATION				MACHINE				PROGRAMMING LANGUAGE		
	AFHRL	NAMRL	FT. RUCKER	FT. KNOX	APPLE	TERAK	PDP 11	ANALOGUE APPARATUS	PASCAL	BASIC	FORTRAN
NON-COGNITIVE											
Activities Interest Inventory	✓						✓				✓
PSYCHOMOTOR											
Two-Handed Coordination	✓							✓			
Complex Coord./Stick & Rudder[4]	✓							✓			
Complex Coordination[5]		✓						✓			
Tank Video Game[6]					✓				✓		
One-Dimensional Compensatory Tracking[7]		✓					✓				
Critical Tracking[7]		✓					✓				
Two-Dimensional Compensatory Tracking			✓		✓					✓	
Kinesthetic Memory			✓		✓					✓	
Helicopter Simulator			✓					✓			
Tank Turret Simulator				✓				✓			
Perceptronics Simulator				✓				✓			
Gunner Tracking Task (using the Willey "Burst-on-Target" Simulator)				✓				✓			
Target Acquisition Task (using the Willey "Burst-on-Target" Simulator)				✓				✓			

[4] AFHRL is currently adapting the Complex Coordination (using two hands) to the PDP 11.

[5] NAMRL is currently adapting this to an Apple computer with joy stick and foot pedals.

[6] Developed under contract with ARI; work being carried out at Pensacola.

[7] These measures administered under NAMRL contract at the Aviation Research Laboratory in Illinois.

COMPUTERIZED MEASURES OBSERVED DURING SITE VISITS FOR ARI PROJECT A

PREDICTOR	LOCATION					MACHINE					PROGRAMMIN LANGUAGE		
	AFHRL	NAMPL	FT. RUCKER	FT. KNOX		APPLE	TERAK	PDP11	ANALOGUE APPARATUS		PASCAL	BASIC	FORTRAN
PSYCHOMOTOR (CONT.)													
Fire Control Computer Task[8] (Using the Chrysler Corp. Fire Control Combat Simulator)				✓					✓				
Round Sensing Task[8] (Using several different pieces of equipment including T-scope, 3 projectors, Allen Device, etc.)				✓					✓				
Computerized Target Engagement (also using 35 mm film, slides, and video equipment)				✓		✓						✓	
Psychomotor Tracking Task				✓		✓						✓	

---

[8] These measures may be more appropriately categorized elsewhere, e.g., Perceptual or Information Processing (Figure memory) for the Round Sensing Task, but have been placed here due to the type of equipment required.